TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi predictiu a 3 mesos de l'índex borsari *Eurostoxx 50* mitjançant tècniques d'aprenentatge automàtic

Autor: Clara Carner Marsal

Director: Ernest Pons Fanals

Departament: Econometria, Estadística i Economia Aplicada

Convocatòria: Juny 2024

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM I PARAULES CLAU

En un context on la cultura de la immediatesa està cada cop més arrelada en la societat i la promesa d'inversions amb guanys instantanis ha guanyat popularitat, aquest treball proposa una alternativa d'inversió contrària a aquesta tendència, enfocada al mitjà termini. Utilitzant la intel·ligència artificial, s'avaluarà el rendiment de models d'aprenentatge automàtic en la predicció a 3 mesos de la taxa de variació de l'índex borsari *Eurostoxx 50* mitjançant tres tècniques d'aprenentatge automàtic: *k-nearest neighbors (KNN)*, *random forest* i *gradient boosting machine (GBM)*. Per modelitzar les dades, aquestes seran preprocessades i s'hi afegiran variables macroeconòmiques per obtenir una base de dades òptima. Dels models resultants, s'escollirà el millor per a cada tècnica en base a una mètrica d'error, en concret l'arrel de l'error quadràtic mig *(RMSE)*. D'entre aquests, es determinarà el model finalista per identificar quina tècnica i quin model específic funcionen millor per a prediccions a 3 mesos, un horitzó temporal menys comú en inversions.

Paraules Clau: Aprenentatge automàtic, predicció financera, *k-nearest neighbors, random forest, gradient boosting machine, Eurostoxx 50*, Mètriques d'error, inversió, fons indexats, cultura de la immediatesa.

ABSTRACT AND KEYWORDS

As the culture of immediacy becomes increasingly entrenched in society and the promise of investments with instant returns gains popularity, this project proposes an alternative investment strategy that counters this trend by focusing on the medium term. The performance of machine learning models in predicting the 3-month rate of change of the *Eurostoxx 50* stock index will be evaluated using three machine learning techniques: k-*nearest neighbors (KNN), random forest*, and *gradient boosting machine (GBM)*. To model the data, preprocessing will be conducted, and macroeconomic variables will be added to obtain an optimal dataset. The best model for each technique will be chosen based on the root mean square error *(RMSE)*. Finally, the most effective technique and specific model for 3-month forecasts, a less common time horizon in investment, will be determined.

Keywords: *Machine learning*, financial prediction, k-nearest neighbors, random forest, gradient boosting machine, *Eurostoxx 50*, *RMSE*, investment, indexed funds, culture of immediacy.

AGRAÏMENTS

En primer lloc vull agrair a l'Ernest Pons, tutor del treball, el fet d'estar sempre disposat a ajudar, per la dedicació, per l'aprenentatge i sobretot per els ànims i la motivació.

Als meus amics per tot el suport.

A la meva família per tot l'acompanyament, paciència, confiança i per ensenyar-me a seguir endavant sempre. I a la meva parella per escoltar-me quan ho necessitava i estar al meu costat, en els moments bons i en els moments més difícils.

CLASSIFICACIÓ AMS

62P20: Applications to economics

62P05: Applications to actuarial sciences and financial mathematics

62M20: Prediction

68T05: Learning and adaptive systems

68Txx: Artificial intelligence

91B84: Economic time series analysis

91G70: Statistical methods;

ÍNDEX DE CONTINGUTS

ÍNDEX DE FIGURES

ÍNDEX DE TAULES

1 INTRODUCCIÓ

En els últims anys, especialment entre els més joves, s'ha observat una marcada cultura de la immediatesa, que comporta la necessitat de tenir-ho tot de manera instantània i sense esforç. Remuntant als seus orígens, les arrels d'aquest fenomen es troben en els primers passos de la globalització, que van derivar en el desenvolupament del comerç internacional, la revolució industrial i, posteriorment, el naixement d'internet. Aquest conjunt de factors ha estès l'economia més enllà de les fronteres, posant a l'abast de qualsevol una àmplia gamma de productes i serveis.

El fenomen d'aquesta cultura de no saber esperar ha evolucionat contínuament i de manera exponencial, gràcies en gran part al desenvolupament d'internet i, sobretot, de les xarxes socials. Per una banda, internet i les tecnologies han permès una comunicació instantània, facilitant l'accés immediat a la informació i als recursos desitjats. Les xarxes socials, al seu torn, han jugat un paper crucial en la difusió de la informació i en la creació de connexions entre individus.

La cultura de la immediatesa té un impacte profund tant en aspectes materials com intangibles de la vida. Un exemple en qüestions materials és l'augment notable de l'ús d'aplicacions que faciliten el lliurament de menjar a domicili. Amb l'acte d'obrir el telèfon mòbil i seleccionar el menjar desitjat, es pot gaudir d'aquest en menys de 30 minuts. Aquestes pràctiques modelen la societat i amplifiquen aquesta necessitat d'immediatesa.

Un altre àmbit on aquesta cultura de la immediatesa es pot veure reflectida és el sector financer, on hi ha una creixent tendència a buscar guanys immediats sense esforç. Aquesta mentalitat es fa evident amb l'augment d'oportunitats d'inversió que prometen rendiments ràpids i fàcils, sovint sense avaluar els riscos associats, com ara les criptomonedes. Aquesta realitat es pot comprovar fàcilment preguntant a persones del nostre entorn quants han fet alguna inversió, i quants d'ells ho han fet en criptomonedes, sovint sense tenir coneixements financers previs.

És en aquest context que aquest treball proposa donar veu als fons indexats, vehicles d'inversió que segueixen l'evolució dels índexs borsaris. S'escullen aquests vehicles d'inversió perquè, donades les seves característiques, són vistos com una opció d'inversió a mitjà o llarg termini, però

mai a curt termini. Així, en lloc de buscar una compensació immediata, s'explora la possibilitat d'invertir amb un horitzó temporal de mitjà termini, com podria ser de tres mesos, rebutjant així la cultura de la immediatesa.

L'objectiu d'aquest treball és analitzar un índex borsari específic anomenat *Eurostoxx 50* i, utilitzant la intel·ligència artificial, concretament la branca de l'aprenentatge automàtic, desenvolupar diversos models predictius per a anticipar el seu rendiment. L'objectiu final és poder predir, amb un horitzó temporal de 3 mesos (90 dies concretament), la taxa de variació, també coneguda com la rendibilitat esperada, d'aquest índex.

L'elecció d'un horitzó temporal a 90 dies (3 mesos) es fonamenta en dos arguments principals. En primer lloc, aquest període es considera un punt entre mig òptim, entre la previsió a curt termini (molt volàtil) i la previsió a llarg termini (més difícil de modelar amb precisió). Un termini intermedi com el de 90 dies permet capturar dinàmiques de mercat significatives, mantenint alhora una precisió raonable en les prediccions. En segon lloc, l'elecció d'un horitzó temporal mitjà resulta interessant pel fet que la major part de la literatura acadèmica es centra en la predicció a curt i a llarg termini. L'ús d'un termini de 90 dies proporciona una perspectiva menys coneguda I estudiada, oferint així noves visions sobre el comportament dels mercats en un termini que no és ni massa breu ni massa extens.

Els models que s'usaran per a la predicció seran tres, *k-nearest neighbors (KNN), random forest i gradient boosting machines (GBM).* Cal destacar que tant en la decisió d'elecció del vehícle d'inversió (fons indexat) i de que les tècniques utilitzades per predir siguin d'intel·ligència artificial, hi juga un gran paper l'interès personal. En quant a l'elecció concreta de l'índex borsari "*Eurostoxx 50*", és deriva en un raó geogràfica i de importància, ja que és el més representatiu d'Europa i també un dels més coneguts.

Cal destacar que en un principi, aquest treball s'havia estructurat de tal manera que l'objectiu no era només predir un índex borsari, sinó també un altre anomenat "S&P 500" i fer la comparació entre aquests dos. També, s'havia plantejat utilitzar una quarta tècnica d'aprenentatge automàtic, concretament les xarxes neuronals. Però, deguda l'extensió del treball un cop realitzada tota la feina corresponent al primer índex i als tres primers models, es va veure que era necessari replantejar el treball i centrar-se només en un índex i en els tres primers models.

En quant a l'estructura d'aquest treball, en primer lloc es presenta el marc teòric, el qual està dividit en dos subapartats principals. El primer, "Els Índexs Borsaris", descriu els mercats financers i els actius financers, proporciona una introducció als índexs borsaris i concretament a l'índex *Eurostoxx 50*. També, revisa la literatura existent sobre l'anàlisi dels índexs borsaris i les variables relacionades. La segona secció, "Tècniques de *machine learning*", inclou una introducció al *machine learning* i descriu els tres mètodes que s'utilitzaran: *KNN, Boscos Aleatoris i gradient boosting machine.*

El tercer apartat, que és la metodologia, descriu detalladament la metodologia emprada en el treball. Aquesta inclou la descripció general de la metodologia, l'anàlisi exploratori de les dades i el seu preprocessament, la selecció i creació de variables per a la construcció de models predictius, aspectes a tenir en compte abans de la modelització com la validació creuada, la selecció de combinacions de variables, les mètriques d'error i la relació entre el paràmetre *ntree* i el concepte de *overfitting*. Finalment, la modelització de models predictius amb les tres tècniques.

El quart apartat, son els resultats, que tal com diu el títol, presenta els resultats obtinguts dels models predictius implementats. Aquest apartat inclou els resultats dels models escollits com a finalistes per a cada mètode i la selecció del millor model per predir el comportament de *l'Eurostoxx 50*. En el cinquè apartat s'hi expliquen les conclusions, des de una perspectiva econòmica i estadística. En el sisè apartat, s'hi presenta la discussió, on es detallen les dificultats trobades durant l'elaboració del treball, els aspectes que es modificarien si es tornés a començar i es proposen línies d'investigació futures per a aquells interessats en continuar l'estudi. Finalment, els últims apartats presenten la bibliografia, el codi (accessible a través d'un enllaç) i l'annex.

2 MARC TEÒRIC

2.1 Introducció

A continuació, s'explicarà el marc teòric d'aquest treball, dividit de manera diferenciada en dos àmbits. Un àmbit es centra en el mercat financer, concretament en els índexs borsaris, i l'altre àmbit tracta sobre el *machine learning*, amb èmfasi en tres tècniques específiques. És important explicar aquests conceptes i d'altres també relacionats per tal de poder entendre bé la base teòrica que sustenta aquest treball.

Cal destacar que al llarg del treball, es farà referència sovint a les tècniques *k-nearest neighbors* (*KNN*), *random forest* i *gradient boosting machine (GBM)* i per simplificar, en moltes situacions es farà us de les seves abreviacions, especialment per al *KNN* i el *GBM*.

En primer lloc, és essencial saber què són els índexs borsaris, especialment l'*Eurostoxx 50*, ja que és l'objectiu principal de les prediccions. Entendre la seva composició i funcionament ajudarà a identificar les variables que poden estar relacionades amb aquest, clau per a construir els models. En segon lloc, cal saber què és l'aprenentatge automàtic i com funcionen les tres tècniques seleccionades: *KNN*, *random forest* i *GBM*. Aquest coneixement és bàsic per a poder construir models de manera adequada i utilitzar aquestes tècniques de la millor manera possible.

2.2 Els índex borsaris

2.2.1 Els mercats financers i els actius financers

Per comprendre els índexs borsaris, primer és fonamental familiaritzar-se amb els conceptes bàsics dels actius i els mercats financers. Aquesta base permetrà entendre com es formen els índexs i quin paper juguen en l'economia.

Un actiu es defineix com qualsevol cosa que pot ser intercanviada, ja sigui un producte, servei, o qualsevol altre bé amb valor de canvi. Els actius es classifiquen en dues categories principals: actius no financers i actius financers. Dins dels actius no financers, es distingeixen els actius tangibles, també coneguts com a actius reals, que tenen una propietat física concreta (per exemple, màquines, edificis o vehicles) i els actius intangibles, que es basen en el dret a obtenir guanys o quantitats monetàries futures (per exemple, patents o drets de propietat intel·lectual).

Els actius financers es poden classificar segons diferents criteris, un dels quals és el tipus de flux que prometen tornar a l'inversor, dividint-se així en actius de renda fixa i actius de renda variable. La renda fixa inclou actius que prometen fluxos de caixa predeterminats, com ara bons, mentre que la renda variable inclou actius amb fluxos de caixa indeterminats, com ara accions. Aquests actius es negocien en les borses de valors. A més, els actius financers es classifiquen segons el termini de venciment: actius de curt termini (venciment en menys d'un any) i actius de llarg termini (durada superior a un any). Dins dels actius financers també es troben diferents tipus com accions, bons, derivats i índexs borsaris.

Un mercat és un lloc, físic o virtual, on s'intercanvien actius, generalment a canvi de diners. Hi ha dos tipus principals de mercats: el mercat de productes (intercanvi de béns i serveis) i el mercat de factors (intercanvi de treball i capital). Dins del mercat de factors, es troba el mercat financer, on es negocien actius financers. Els mercats financers tenen tres funcions principals: posar en contacte els emissors d'actius financers amb els inversors, servir com a mecanisme de formació i fixació de preus gràcies a la interacció entre compradors i venedors, i proporcionar liquiditat als actius, facilitant-ne la venda de manera ràpida i eficient. Els mercats financers posseeixen diverses característiques que garanteixen el seu bon funcionament, les quals es resumeixen en la següent taula:

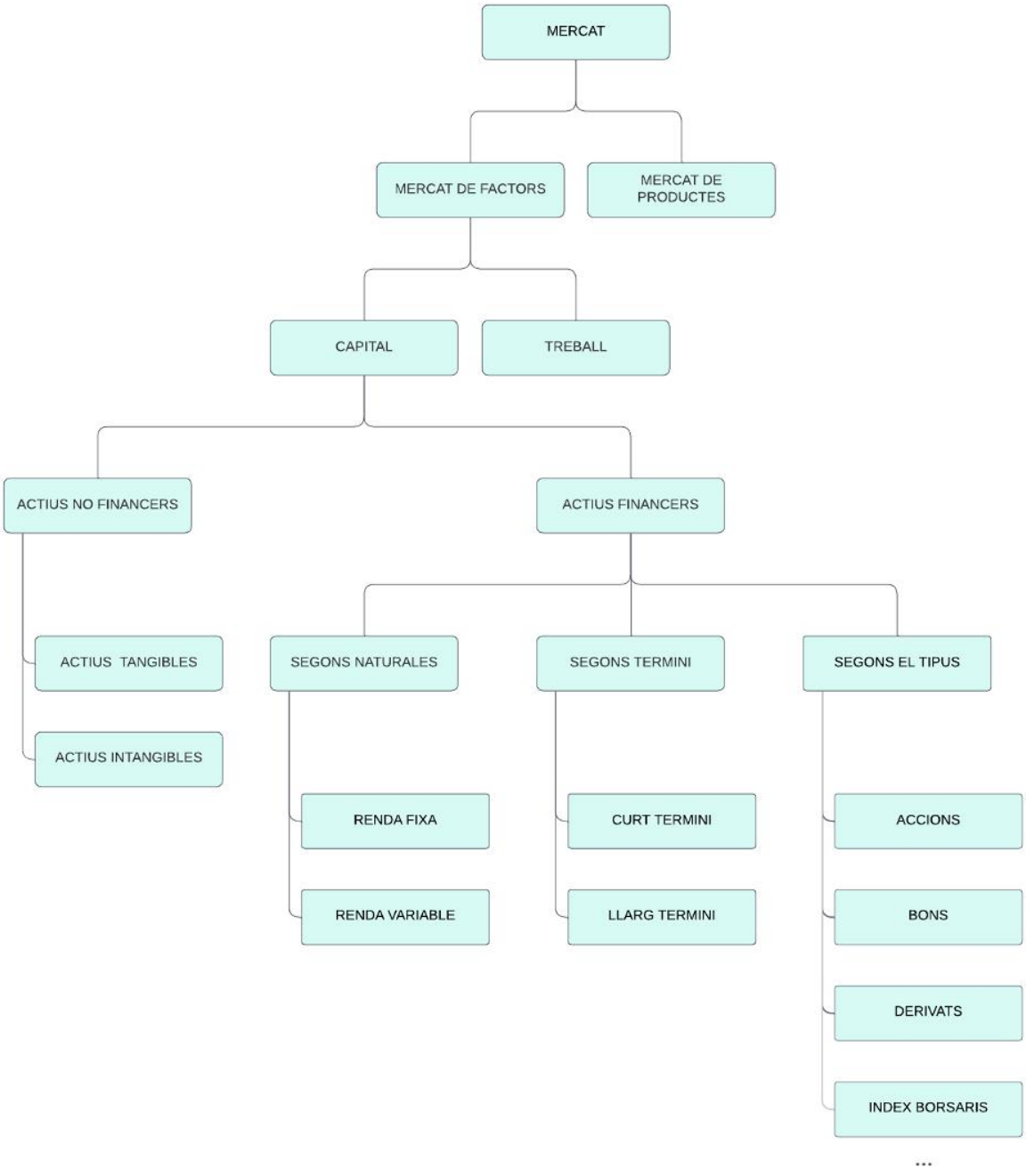
Figura 2.2.1 Definició de les característiques dels mercats financers

Variable	Definició
Transparència	És essencial que la informació sobre els actius sigui accessible i clara per a tots els participants del mercat. Això inclou la disponibilitat de dades actualitzades sobre els preus, volums de transacció i altres informacions rellevants per a la presa de decisions.
Llibertat	Fa referència a la no limitació per a l'accés de compradors i venedors, permetent una lliure formació de preus sense influències externes. Això implica l'absència de barreres d'entrada o sortida i la possibilitat de negociar actius en les quantitats desitjades.
Profunditat	Un mercat és profund quan pot absorbir grans ordres de compra o venda sense afectar significativament el preu dels actius. Això significa que hi ha suficients compradors i venedors per mantenir l'estabilitat dels preus davant de grans volums de transaccions.
Amplitud	Fa referència a la varietat d'actius que es negocien en un mercat. Un mercat ampli ofereix una àmplia gamma d'actius financers diferents, el que permet als inversors diversificar les seves carteres i reduir els riscos.
Flexibilitat	És la capacitat del mercat per reaccionar ràpidament davant canvis en els preus dels actius. Un mercat flexible ajusta els preus ràpidament quan es detecten desajustos, permetent que els preus reflecteixin ràpidament la informació nova.

Font: Elaboració pròpia a partir dels apunts de l'assignatura "instruments i mercats financers"

Un cop establertes les bases conceptuals, el següent esquema mostra de manera visual on es classifiquen els índexs borsaris dins de l'ecosistema dels actius i mercats. L'esquema presenta com els índexs borsaris, com a part dels actius financers de renda variable, s'integren en els mercats de capital, que al seu torn formen part del mercat de factors:

Figura 2.1 Mapa conceptual mercats i índex borsaris



Font: Elaboració pròpia a partir dels apunts de l'assignatura "instruments i mercats financers"

2.2.2 Índex borsaris

2.2.2.1 Introducció

Els índexs borsaris són un tipus específic d'actiu financer de renda variable que representa el valor conjunt d'un grup d'accions seleccionades. Aquests índexs proporcionen una mesura del rendiment del mercat o d'un segment del mercat, i són utilitzats com a referència per avaluar la rendibilitat d'inversions específiques. Un exemple d'índex borsari és l'*Eurostoxx 50*, que inclou les 50 empreses més grans de la zona euro. Els índexs borsaris no es negocien directament com les accions o els bons, però es poden comprar i vendre a través de productes financers derivats com fons indexats, els quals tracten de replicar el rendiment dels índexs. Aquests productes es negocien en les borses de valors.

Com ja s'ha comentat prèviament, aquest treball es centra específicament en els índexs borsaris, concretament en el *Eurostoxx 50.* Els índexs borsaris serveixen com a baròmetres del rendiment del mercat, proporcionant als inversors una eina per avaluar el comportament general del mercat o d'un sector específic. Aquests, poden variar en la seva composició i els sectors que representen, oferint una visió general de l'economia o d'àrees específiques. El factor més important en la construcció d'un índex és la contribució del pes de cadascun dels seus components.

Hi ha tres mètodes principals per fer-ho: la ponderació per preu (*price weighted*), la ponderació per valor (*value weight o market-cap weight*) i la ponderació igualitària (*equal weight*). (Per a més informació sobre aquests tres mètodes, visitar: https://www.thebalancemoney.com/different-types-of-weighted-indexes-1214780)

2.2.2.2 Eurostoxx 50

L'Eurostoxx 50 és l'índex borsari de referència a Europa, que agrupa les 50 empreses més grans de vuit països de l'Eurozona, conegudes com a "blue chips" [1]. Aquestes empreses, que pertanyen

[1] Les empreses *blue chips* es defineixen per ser companyies grans, ben establertes i financerament sòlides. Aquestes ofereixen rendiments estables i fiables als seus inversors.

a 19 sectors diferents, tenen el sector bancari com a principal contribuent en la seva ponderació. L'índex es va establir el 31 de desembre de 1991 amb una base de 1.000 punts, tot i que no es va crear formalment fins al 1998. Va ser desenvolupat conjuntament per *Dow Jones & Company, Deutsche Börse* i *Swiss Exchange*, sota l'aliança de *Stoxx Limited.*

L'Eurostoxx 50 es calcula en funció de la capitalització borsària ajustada pel *free float*[2], cosa que significa que les empreses amb més valor i liquiditat tenen més influència en l'índex. Tot i això, cap empresa pot tenir més d'un 10% de pes. La composició de l'índex es revisa anualment al setembre per assegurar que continua reflectint el mercat europeu. L'índex *Eurostoxx 50* es calcula utilitzant la fórmula de *Laspeyres*, que permet ajustar els canvis en les quantitats i els preus dels components de l'índex. La fórmula en concret és:

$$\text{Index}_t = \frac{\sum_{i=1}^{n}(p_{it} \cdot s_{it} \cdot ff_{it} \cdot cf_{it} \cdot x_{it})}{D_t}$$

On:

- $\boldsymbol{t}$: Instant de temps específic en què es calcula l'índex
- $\boldsymbol{p_{it}}$: Preu de l'acció de la companyia i en el moment t.
- $\boldsymbol{s_{it}}$: Nombre d'accions de la companyia i en el moment t.
- $\boldsymbol{ff_{it}}$: Factor de flotació lliure de la companyia i en el moment t.
- $\boldsymbol{cf_{it}}$: Factor de ponderació de la companyia i en el moment t.
- $\boldsymbol{x_{it}}$: Tipus de canvi de la moneda local a la moneda de l'índex per a la companyia i en el moment t.
- $\boldsymbol{Dt}$: Divisor de l'índex en el moment t.

Quan una empresa emet més accions, la seva capitalització borsària augmenta, cosa que teòricament hauria d'augmentar l'índex. No obstant això, per evitar canvis abruptes i mantenir la continuïtat de l'índex, es fa un ajust en el divisor, sent aquest:

[2] "El free float" es coneix com a la part de les accions d'una empresa que cotitzen lliurement al mercat, disponibles per ser comprades i venudes pel públic sense restriccions.

$$D_{t+1} = D_t \times \frac{\sum_{i=1}^{n}(p_{it} \cdot s_{it} \cdot ff_{it} \cdot cf_{it} \cdot x_{it}) \pm \Delta MC_{t+1}}{\sum_{i=1}^{n}(p_{it} \cdot s_{it} \cdot ff_{it} \cdot cf_{it} \cdot x_{it})}$$

On, $\boldsymbol{\Delta\, MC_{t+1}}$ és la diferència entre la capitalització borsària [3] ajustada i la capitalització borsària de tancament de l'índex després d'un esdeveniment corporatiu.

Aquest ajust assegura que l'índex només reflecteixi els canvis reals en els preus de les accions, no en el nombre d'accions. Així, l'índex no augmenta simplement perquè una empresa emeti més accions; només augmenta si els preus de les accions realment pugen. L'ajust del divisor serveix per garantir que l'índex sigui una representació precisa dels moviments del mercat i que no es vegi determinat per accions corporatives, com per exemple, un split[4].

Per a més detalls sobre el càlcul de l'index, es pot consultar el següent link: https://en.wikipedia.org/wiki/EURO_STOXX_50

Actualment, des de la darrera actualització (setembre de 2023), els països que compten amb empreses a l'*Eurostoxx 50* són els següents:

[3] Valor total de mercat d'una companyia. Es calcula multiplicant el preu de l'acció, en un moment determinat, pel nombre total d'accions.

[4] Acció corporativa en la qual una empresa divideix les seves accions existents en múltiples accions per augmentar la liquiditat

Taula 3.2 Nombre d'empreses per país de l'Eurostoxx 50

Ranquing	Pais	Empreses
1	França	17
2	Alemanya	14
3	Països Baixos	6
4	Itàlia	5
5	Espanya	4
6	Finlàndia	2
7	Bèlgica	1
7	Irlanda	1
	Total	**50**

Font: Elaboració pròpia a partir de dades obtingudes a wikipedia

2.2.3 Literatura sobre l'anàlisi dels índexs borsaris i les variables relacionades

L'estudi dels índexs borsaris, incloent aquest el seu anàlisi i predicció, ha estat àmpliament tractat en la literatura acadèmica. Sovint, aquesta recerca s'ha enfocat en identificar les variables que poden afectar els moviments dels mercats financers.

Un dels models més comuns utilitzats històricament per predir els índexs borsaris ha sigut el model autoregressiu integrat de mitjana mòbil (ARIMA). Aquest model va ser desenvolupat per Box i Jenkins (1976) i ha sigut molt utilitzat amb la finalitat de predir els índexs borsaris a partir de dades passades (Zamorano Cid & Giménez Fernández, s.d., p. 12).

Tot i que aquests models han demostrat en varies ocasions la seva eficàcia, com ja s'ha comentat en aquest treball només s'estudiaran i s'utilitzaran tècniques de *machine learning*.

Respecte les tècniques de *machine learning* en aquest àmbit, el canvi i l'utilització cap aquestes ha augmentat de manera significativa en els darrers anys, degut a la seva capacitat per manejar grans quantitats de dades i identificar patrons complexos que poden no ser evidents amb els models tradicionals. Aquests, presenten avantatges significatius en termes de precisió i

flexibilitat, permetent que dades no lineals i heterogènies puguin ser modelades de manera òptima.

En el desenvolupament dels models predictius en les tres tècniques utilitzades, *KNN, GBM i random forest*, es tindran en compte dues variables macroeconòmiques que s'han demostrat importants i significatives en la literatura: el PIB i la inflació.

El PIB és un indicador de la salut econòmica general d'un país. Un augment en el PIB sol estar correlacionat amb un increment en els índexs borsaris, ja que implica una economia en expansió i una major confiança dels inversors. Diversos estudis han confirmat la importància del PIB com a variable explicativa dels moviments dels índexs borsaris. Glenn (2009) va demostrar la importància del PIB en models predictius utilitzant el PIB dels Estats Units per predir els moviments del S&P 500 (Zamorano Cid & Giménez Fernández, s.d., p. 12). Khan, M. S. & Sharif (2012) també van trobar una correlació positiva entre el creixement del PIB i l'augment dels índexs borsaris en diversos països (Zamorano Cid & Giménez Fernández, s.d., p. 20).

La segona variable és la inflació, que afecta el poder adquisitiu i pot influir en les decisions d'inversió. Una inflació alta tendeix a reduir la confiança dels inversors, fet que pot tenir un efecte negatiu en els índexs borsaris. L'impacte de la inflació en els índexs borsaris ha estat àmpliament estudiat, demostrant la seva importància com a variable predictiva. Khan, M. S. & Sharif (2012) van analitzar com la inflació afecta els índexs borsaris en diversos països, concloent que hi ha una relació negativa significativa entre la inflació i els índexs borsaris (Zamorano Cid & Giménez Fernández, s.d., p. 20).

2.3 Tècniques de machine learning

2.3.1 Introducció al machine learning

El *machine learning*, o aprenentatge automàtic, és una subcategoria de la intel·ligència artificial que permet als algoritmes descobrir patrons recurrents (*patterns*) en conjunts de dades. Aquestes dades poden ser números, paraules, imatges, etc. Qualsevol informació que es pugui emmagatzemar digitalment pot servir com a dada per al *machine learning*. Un cop identificats els patrons, aquests algoritmes són capaços d'aprendre i millorar el seu rendiment de manera autònoma, amb l'objectiu de realitzar tasques específiques amb major eficiència i precisió. Aquesta branca de la IA se centra en l'ús de dades i algoritmes per imitar l'aprenentatge humà, permetent que les màquines millorin amb el temps i siguin cada cop més precises en la realització de prediccions o classificacions. Una vegada entrenat, l'algoritme és capaç de trobar patrons en noves dades. Dit de manera més simple, el *machine learning* consisteix en mostrar un gran volum de dades a una màquina perquè pugui aprendre a fer prediccions, trobar patrons o classificar dades.

Els tres tipus de *machine learning* són supervisat, no supervisat i d'aprenentatge per reforç. En l'aprenentatge supervisat, la màquina aprèn a partir de dades etiquetades que proporcionen exemples de la sortida desitjada. Aquest és el tipus de *machine learning* més utilitzat en el món empresarial, amb aplicacions com la previsió de vendes, l'optimització d'inventaris i la detecció de fraus. L'aprenentatge no supervisat treballa amb dades no etiquetades i busca descobrir patrons ocults en les dades. Aquest tipus d'aprenentatge és molt útil per identificar patrons i prendre decisions basades en les dades, amb aplicacions comunes com el clustering i la identificació d'associacions en les dades. L'aprenentatge per reforç és el tipus de *machine learning* més semblant a l'aprenentatge humà, on l'algoritme aprèn a través de la interacció amb l'entorn i rebent recompenses positives o negatives. Algunes aplicacions inclouen l'entrenament de vehicles autònoms per conduir, la regulació dinàmica de semàfors per reduir embussos i l'entrenament de robots utilitzant imatges de vídeo. (Coursera, 2023)

Per predir l'índex borsari, aquest treball utilitzarà tres tècniques de *machine learning* supervisat: *KNN, random forest i GBM.* Aquestes tècniques han estat seleccionades per la seva capacitat de gestionar grans quantitats de dades històriques i identificar patrons complexos que poden no ser evidents amb els models tradicionals. És important destacar que aqusts tres algoritmes es poden

utilitzar tant per a la classificació com per a la predicció. No obstant això, en aquest treball només es farà ús de la seva capacitat de predicció. Tant el *random forest* com el *GBM* són mètodes d'ensamblatge que combinen les prediccions de múltiples models individuals per millorar la precisió i la robustesa del resultat final, a diferència del *KNN*, que és un model individual.

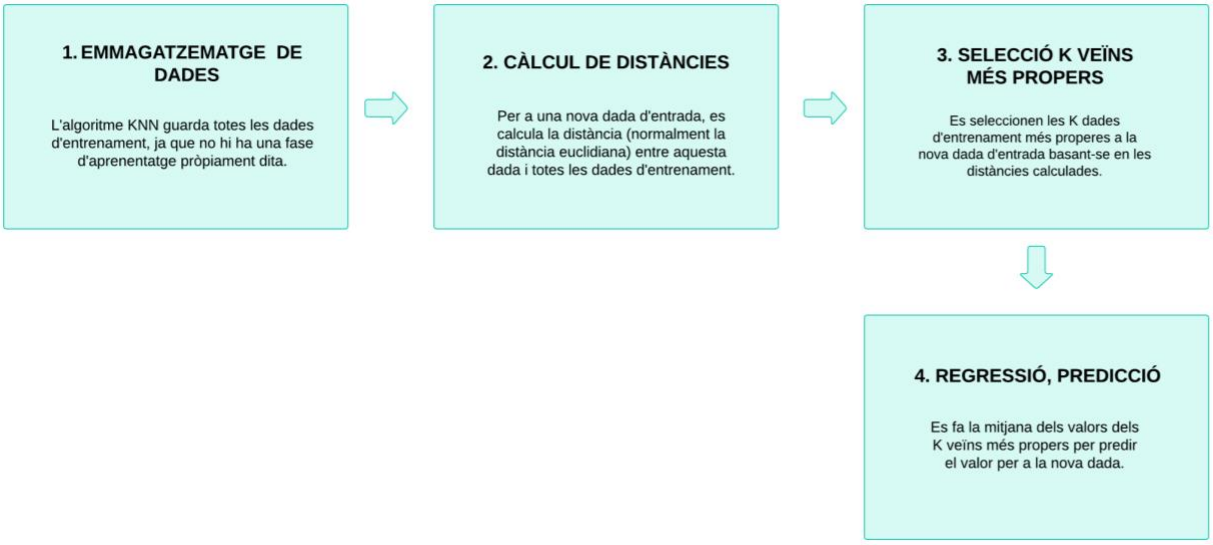
2.3.2 Mètodes de machine learning

2.3.2.1 KNN

El model *KNN* (veïns més propers) és un dels algoritmes de classificació i regressió més senzills i més utilitzats en l'aprenentatge automàtic. El seu funcionament es basa en trobar els k exemples d'entrenament més propers a un nou punt de dades per fer una predicció. Cal destacar que realment aquest no crea un model, sinó que utilitza les dades d'entrenament per tal de predir (GeeksforGeeks, 2017).

A continuació s'explicarà de manera esquemàtica el recorregut que fa l'algoritme *KNN*:

Figura 2.2 Esquema de les etapes de l'algoritme KNN

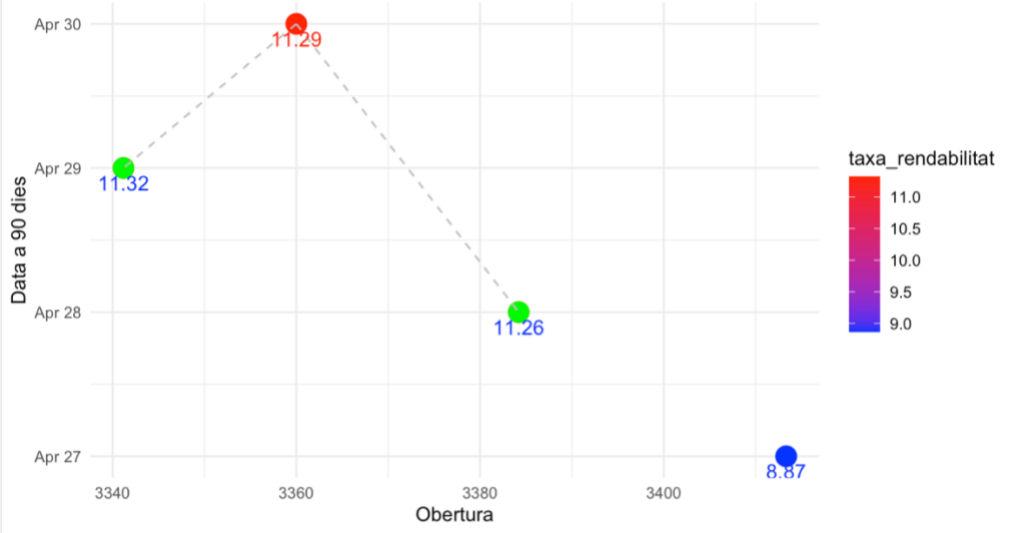


Font: Elaboració pròpia a partir de la informació de "GeeksforGeeks"

En el següent gràfic, es presenta el model *KNN* de manera visual, utilitzant una versió simplificada de la base de dades que s'utilitzarà posteriorment. L'objectiu és facilitar la comprensió del procés que segueix l'algoritme. S'han seleccionat dues variables explicatives, "Obertura" i "Data a 90 dies", i la variable objectiu, anomenada "taxa_rendabilitat". Els punts blaus i verds representen les dades d'entrenament, amb el color indicant la "taxa de rendibilitat". El punt vermell representa la nova dada per a la qual es vol predir la rendibilitat, i els punts verds són els veïns més propers seleccionats pel model.

Les línies grises discontínues connecten la nova dada amb els seus veïns més propers, mostrant visualment les distàncies utilitzades pel model per fer la predicció. Les etiquetes sota els punts indiquen els valors de la rendibilitat, sent la vermella la predicció del model per a la nova dada, obtinguda a partir de la mitjana de les dues verdes. Aquest exemple ajuda a entendre el procés que s'utilitzarà en el model complet, el qual inclourà més variables i dimensions, fent que la seva representació gràfica sigui impossible.

Figura 2.3 Exemple visual del funcionament del KNN



Font: Elaboració pròpia a R amb l'algoritme KNN a partir de variables seleccionades de la base de dades

Per obtenir més informació sobre l'estat actual dels coneixements disponibles sobre l'algoritme KNN, es pot consultar l'enllaç següent: https://link-springer-com.sire.ub.edu/chapter/10.1007/978-3-540-39964-3_62

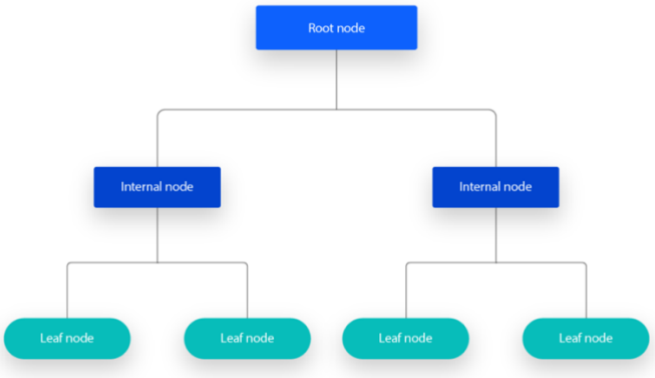
2.3.2.2 Boscos aleatoris

La tècnica de bosc aleatori *(random forest)* és una eina d'aprenentatge automàtic que pot realitzar tant tasques de regressió com de classificació. Aquest algoritme es basa en la creació de múltiples arbres de decisió, cadascun dels quals proporciona una predicció o una classificació (Contributors to Wikimedia projects, 2024).

Per comprendre millor l'algoritme, primer es necessari conèixer què són els arbres de decisió. Aquests tenen una estructura jeràrquica que comença amb un node arrel i es divideix en nodes interns que prenen decisions basades en diverses característiques, acabant en nodes fulla que representen els resultats finals. Els arbres de decisió utilitzen una estratègia de divisió per crear subconjunts homogenis i poden emprar tècniques per evitar el sobreajustament[5], millorant així la simplicitat i precisió del model.

L'estructura general d'un arbre de decisió es representa visualment de la següent manera:

Figura 2.4 Arbre de decisió



Font: IBM , ¿Qué es un árbol de decisión?

En el cas de la regressió, el *random forest* fa la mitjana dels valors predits per cada arbre per obtenir la predicció final. Aquesta metodologia permet generar prediccions robustes i precises, aprofitant la diversitat i la complementarietat dels diferents arbres de decisió creats a partir dels

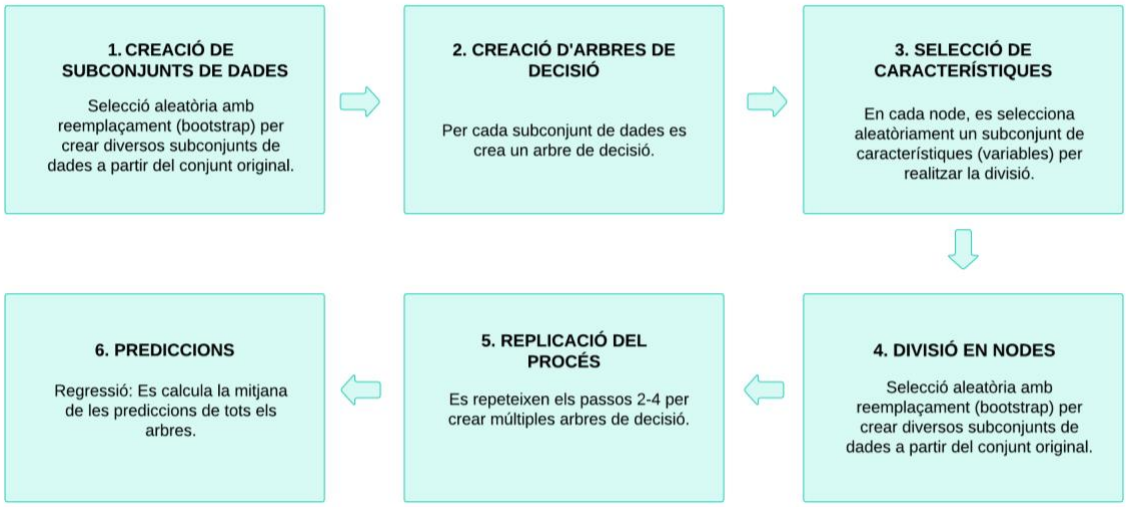
[5] El sobreajustament es produeix quan un model aprèn tant les dades d'entrenament que perd la capacitat de generalitzar a noves dades.

subconjunts de dades. Com molt bé indica el seu nom, el primer pas que es realitza és la creació de diferents conjunts de dades, seleccionades aleatòriament, per crear diferents arbres de decisió, el que acaba sent un bosc, degut a que compren diferents arbres creats amb dades aleatòries.

Tal com indica el seu nom, el procés de creació del *random forest* comença amb la generació de diferents conjunts de dades seleccionades aleatòriament *(random)* per formar diversos arbres de decisió *(forest).* Aquests arbres, combinats, formen un bosc gràcies a l'ús de dades aleatòries per crear múltiples arbres de decisió.

A continuació s'explicarà de manera esquemàtica el recorregut que fa aquest algoritme :

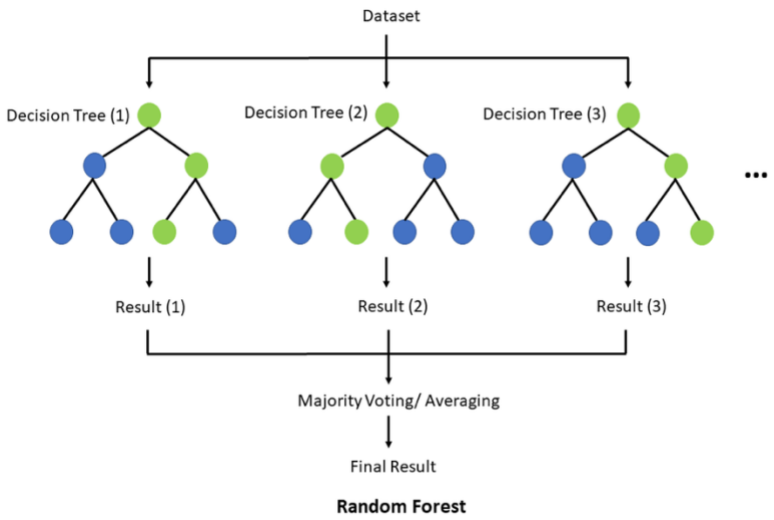
Figura 2.5 Esquema de les etapes de l'algoritme random forest



Font: Elaboració pròpia a partir de la informació de "Wikipedia"

L'estructura general que presenta un bosc aleatori és:

Figura 2.6 Estructura general d'un bosc aleatori



Font: InteractiveChaos

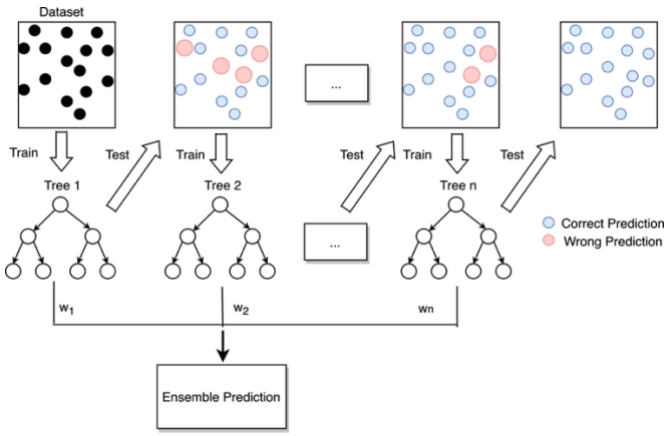
Per obtenir més informació sobre l'estat actual dels coneixements disponibles sobre l'algoritme *random forest*, es pot consultar l'enllaç següent: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

2.3.2.3 Gradient boosting machine

El *gradient boosting machine (GBM)* és un algoritme d'aprenentatge automàtic que es basa en la creació de múltiples arbres de decisió de manera seqüencial, on cada nou arbre intenta corregir els errors comesos pels arbres anteriors. A diferència del *random forest*, que crea arbres de decisió de manera independent i en paral·lel, el *GBM* construeix cada arbre successivament, influenciat pels errors dels arbres anteriors (Amat Rodrigo, 2017).

Una manera gràfica d'entendre l'algoritme seria:

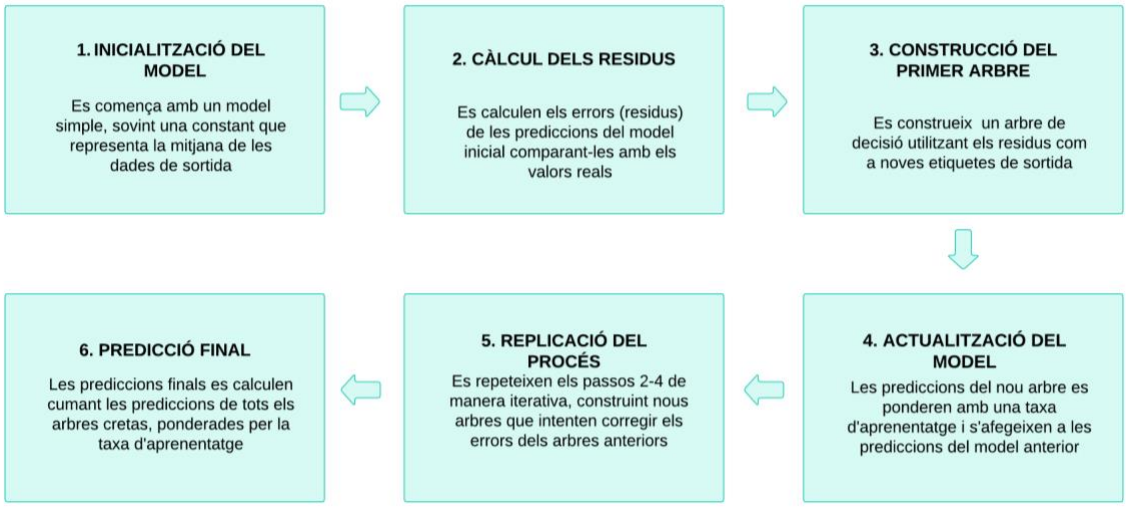
Figura 2.7 Representació visual del procés de l'algoritme GBM



Font: ResearchGate

El recorregut que fa l'algoritme és el següent:

Figura 2.8 Esquema de les etapes de l' algoritme GBM



Font: Elaboració pròpia a partir de la informació de "DataCamp"

Els tres aspectes principals en què el Gradient Boosting es diferencia del random forest són la construcció seqüencial versus la paral·lela, la correcció d'errors i la profunditat dels arbres. En el *random forest*, els arbres es construeixen en paral·lel i de manera independent, utilitzant un subconjunt de dades seleccionat aleatòriament amb reemplaçament *(bootstrapping*[6]*).* En canvi, en el *GBM*, els arbres es construeixen de manera seqüencial, on cada arbre nou intenta corregir els errors dels arbres anteriors, permetent així una millora iterativa del model.

Pel que fa a la correcció d'errors, el *random forest* combina les prediccions de múltiples arbres independents mitjançant la mitjana per regressió, reduint la variància del model i millorant la precisió. En contraposició, el *GBM* es centra en corregir els errors dels arbres anteriors, construint cada arbre nou per predir els residus (errors) de les prediccions anteriors, millorant així la precisió del model amb cada iteració.

Finalment, la profunditat dels arbres és un altre factor diferencial. Els arbres en el *random forest* solen ser més profunds, amb molts nivells, perquè cada arbre és entrenat independentment i intenta ajustar-se bé a les dades de formació. En canvi, els arbres en el *GBM* són més superficials, amb menys nivells, per evitar el sobre ajustament. Això permet que els arbres corregeixin errors específics sense adaptar-se massa a les dades de formació.

[6] Tècnica estadística que implica la selecció aleatòria de subconjunts de dades amb reemplaçament, on cada element seleccionat es torna a posar al conjunt original per poder ser escollit de nou.

Per saber més sobre l'estat actual de la informació disponible sobre l'algoritme, es pot consultar l'enllaç següent. Aquest tracta específicament de XGBoost, una implementació avançada del *GBM*: Informació sobre XGBoost.

3 METODOLOGIA

3.1 Descripció de la metodologia emprada

A continuació, es presenta una visió general de la metodologia seguida, però sense entrar en detall, ja que aquests aspectes es tractaran amb profunditat en els subapartats següents.

La primera fase consisteix en la recopilació de dades històriques de l'índex Euro Stoxx 50. En aquesta, s'extreuen dades de la web "investing.com". Es selecciona un període temporal que inclou dades diàries des del 1 de gener de 2015 fins al 25 de març de 2024. Es selecciona aquest ja que es considera que es un període significatiu per a poder fer l'anàlisi i prediccions d'aquestes.

Un cop recopilades les dades, es porta a terme una anàlisi exploratòria de les dades per analitzar com son aquestes, i per observar si son valides per el seu anàlisi o s'hi han de fer transformacions. Aquesta fase inclou la generació d'estadístiques descriptives

Seguidament, es porta a terme un procés de preprocessament per assegurar la qualitat de les dades utilitzades. Aquest procés inclou la neteja de dades, la imputació de valors perduts, la normalització de noms de variables, i la transformació de dades quan és necessari.

Abans de començar amb la modelització, s'exposen diversos aspectes importants. Es comença explicant l'utilització de la validació creuada i el motiu (assegurar l'aleatorietat dels models i controlar el sobre ajustament). També, s'explica i es determina com es decidiran les variables que s'utilitzaran en cada model (provant totes les combinacions). Per últim, es defineixen les mètriques d'error (per avaluar els models) i es determina quin criteri s'utilitzarà per a definir un model com a millor (minimització de l'arrel de l'error quadràtic mig o *RMSE*).

Durant la fase de modelització, es construeixen i entrenen els models predictius a R utilitzant les 3 tècniques de *machine learning* seleccionades. Els models són ajustats per optimitzar la precisió de les prediccions i concretament, per minimitzar les mètriques d'avaluació, que son els errors. Es construeixen models específics per a cada tècnica: *k-nearest neighbors (KNN)*, utilitzat per la seva simplicitat i efectivitat en problemes de regressió; *Boscos Aleatoris (random forest)*, seleccionat per la seva capacitat de manejar dades amb moltes variables i per reduir la variància de les prediccions; i *gradient boosting machine (GBM)*, utilitzat per la seva capacitat de millorar iterativament les prediccions corregint els errors dels models anteriors. Cal remarcar, que també s'han escollit aquests tres per el coneixement previ que es tenia d'aquests gràcies a la assignatura "Mètodes Estadístics en Mineria de Dades".

La validació dels models es realitza mitjançant tècniques de validació creuada per assegurar que els resultats obtinguts siguin generals i no estiguin afectats pel sobre ajustament. Es calculen diverses mètriques d'error per a cada model, com el *mean square error (MSE)*, el *mean absolute error (MAE)*, el *root mean squared error (RMSE)* i el *mean absolute percentage error (MAPE)*

Finalment, es selecciona el millor model per a cada tècnica basant-se en les mètriques d'error, i posteriorment es determina el model finalista que funcionarà millor per a les prediccions a 3 mesos.

Seguidament doncs, s'entrarà en detall en l'explicació de cada fase realitzada, començant per l'anàlisi de la base de dades i el preprocessament d'aquestes.

3.2 Anàlisi exploratori de les dades i preprocessament d'aquestes

Es considera de gran rellevància entendre correctament la base de dades amb la que es treballa, ja que una interpretació errònia pot comprometre la fiabilitat de l'anàlisi. Així doncs, abans de procedir amb qualsevol tractament o anàlisi, és molt important portar a terme una comprensió detallada de la naturalesa de les dades i les seves variables.

En la següent taula es mostra de manera resumida una visió general de la base de dades inicial:

Figura 3.3.1 Primers registres de la base de dades d'Eurstoxx 50 importada

Base de dades inicial Eurostoxx 50

Fecha	Último	Apertura	Máximo	Mínimo	Vol.	X..var.
25.03.2024	5.019,95	5.031,55	5.039,35	5.013,35		-0,22%
22.03.2024	5.031,15	5.035,13	5.037,74	5.008,90	27,33M	-0,42%
21.03.2024	5.052,31	5.042,32	5.058,91	5.021,14	27,74M	1,04%
20.03.2024	5.000,31	4.991,37	5.009,13	4.980,95	23,45M	-0,15%
19.03.2024	5.007,92	4.979,06	5.007,92	4.976,37	27,40M	0,50%
18.03.2024	4.982,76	4.993,66	5.004,08	4.976,61	24,81M	-0,07%

Font: Elaboració pròpia a partir de la web investing.com

Aquesta base de dades, tal com s'ha mencionat prèviament s'ha descarregat de la següent pàgina web: *https://es.investing.com/indices/eu-stoxx50-historical-data*

Per procedir amb l'anàlisi de dades, en primer lloc es realitzen algunes modificacions bàsiques per assegurar que R interpreti correctament els valors. Per exemple, es determina el punt com a separador decimal i es converteix la variable "Vol." al valor corresponent eliminant la lletra "M". També es canvien els formats de les variables a aquells necessaris amb la finalitat de que cada variable sigui identificada correctament, com ara les dates i els percentatges. A més, es normalitzen els noms de les variables per a alinear-los amb el treball, redefinint-les en català.

Un cop realitzats aquests ajustos bàsics, es procedeix a fer un breu anàlisi descriptiu, per tal de conèixer més les dades. Es començarà per descriure a nivell conceptual les variables.

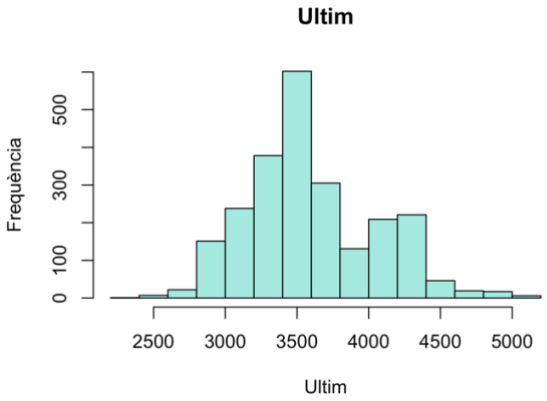
Taula 3.3.1 Descripció de les variables de la base de dades inicial

Variable	Definició
Data	El dia en què es van registrar les dades de l'índex.
Últim	El valor de tancament de l'índex en aquell dia específic. Representa el preu final de l'índex en tancar la jornada borsària.
Obertura	El valor de l'índex a l'inici de la jornada de negociació. Indica el preu amb què es va obrir el mercat aquell dia.
Màxim	El valor més alt que va assolir l'índex durant la jornada de negociació.
Mínim	El valor més baix que va assolir l'índex durant la jornada de negociació.
Volum	Indica el nombre d'accions de les empreses de l'Eurostoxx 50 que es van negociar durant la jornada.
Variació percentual	Mostra el canvi percentual del valor de l'índex respecte al valor de tancament del dia anterior.

Font: Elaboració pròpia

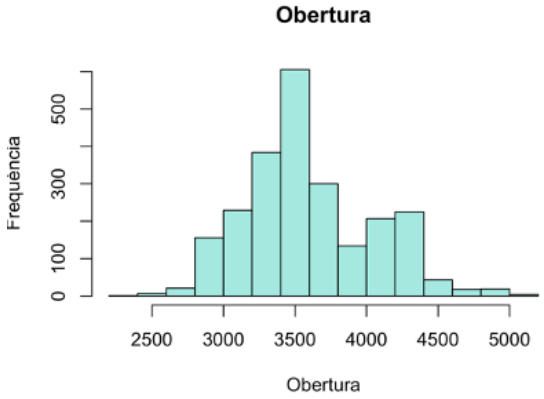
En segon lloc, es procedeix a fer un breu anàlisi descriptiu d'aquestes:

Figura 3.3.3 Distribució de la variable "Ultim"



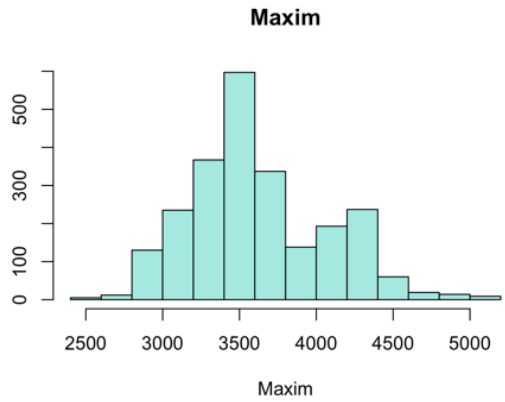
Font: Elaboració pròpia a partir de la BDD importada de investing.com

Figura 3.2 Distribució de la variable "Obertura"



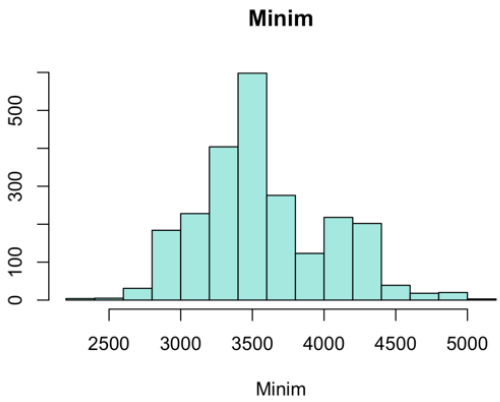
Font: Elaboració pròpia a partir de la BDD importada de investing.com

Figura 3.3.5 Distribució de la variable "Maxim"



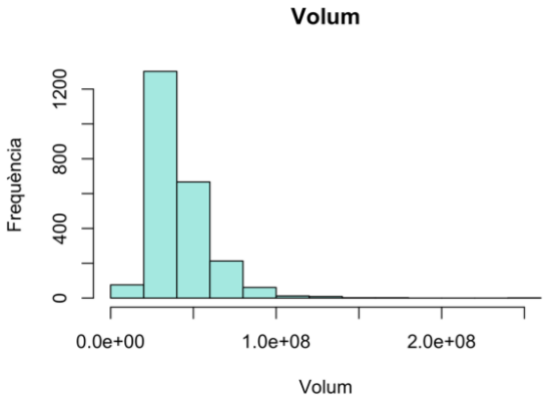
Font: Elaboració pròpia a partir de la BDD importada de investing.com

Figura 3.4 Distribució de la variable "Minim"



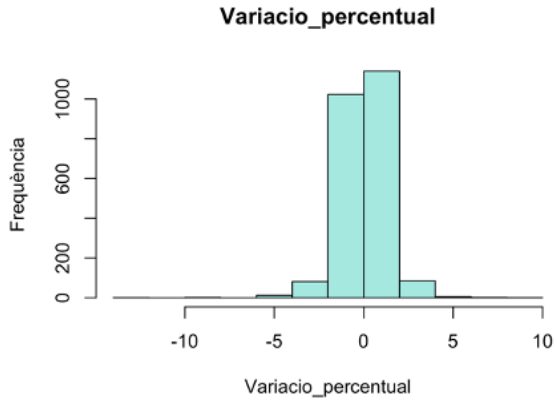
Font: Elaboració pròpia a partir de la BDD importada de investing.com

Figura 3.3.7 Distribució de la variable "Volum"



Font: Elaboració pròpia a partir de la BDD importada de investing.com

Figura 3.6 Distribució de la variable "Variacio_percentual"



Font: Elaboració pròpia a partir de la BDD importada de investing.com

Figura 3.8 Resum estadístic de la base de dades inicial

Resum estadístic de les dades

Data	Ultim	Obertura	Maxim	Minim	Volum	Variacio_percentual
Min. :2015-01-26	Min. :2386	Min. :2389	Min. :2462	Min. :2303	Min. : 11040000	Min. :-12.4000
1st Qu.:2017-05-10	1st Qu.:3305	1st Qu.:3306	1st Qu.:3326	1st Qu.:3279	1st Qu.: 29300000	1st Qu.: -0.5100
Median :2019-08-26	Median :3519	Median :3519	Median :3537	Median :3498	Median : 36860000	Median : 0.0500
Mean :2019-08-25	Mean :3597	Mean :3597	Mean :3620	Mean :3573	Mean : 41261002	Mean : 0.0242
3rd Qu.:2021-12-09	3rd Qu.:3878	3rd Qu.:3874	3rd Qu.:3914	3rd Qu.:3843	3rd Qu.: 48170000	3rd Qu.: 0.6100
Max. :2024-03-25	Max. :5052	Max. :5042	Max. :5059	Max. :5021	Max. :243140000	Max. : 9.2400
NA	NA	NA	NA	NA	NA's :8	NA

Font: Elaboració pròpia a partir de la base de dades importada de investing.com

Després de realitzar un anàlisi detallat de les dades, s'observa que, en general, mostren una estructura coherent. No obstant això, es detecta que la variable "volum" presenta 8 valors faltants

(*missings)*. Amb l'objectiu de preservar la integritat del conjunt de dades i evitar la pèrdua d'informació rellevant, s'opta per aplicar una tècnica d'imputació de dades. Aquest procés es realitza mitjançant l'ús d'una tècnica anomenada *MICE*. Concretament, el paquet de *mice* realitza substitucions mitjançant models predictius. En R, s'ha implementat de la següent manera:

```
imputacio <- mice(data, method = "pmm", m = 5)
```

El mètode *MICE* es pot basar en diferents tècniques, i en aquest treball s'ha escollit el mètode *predictive mean matching (PMM)*. Aquest segueix l'esquema presentat per Van Buuren (2018):

En primer lloc es construeix un model de regressió per predir els valors que falten utilitzant les dades observades. Per exemple, si s'està imputant la variable "Y":

$$Y \sim X1 + X2 + \cdots + Xn \qquad \text{on, X1, ... , Xn son variables del conjunt de dades}$$

Seguidament, utilitzant el model de regressió, es calcula la predicció per a Y per a totes les observacions del conjunt de dades (tant les observacions amb valors observats de "Y", com les observacions amb valors mancants de "Y"). En el següent pas, per a cada valor faltant de "Y" es determinen els valors observats de "Y" que tenen les prediccions més properes a la predicció del valor mancant.

Per últim, d'entre els valors observats més propers, es selecciona un valor aleatòriament per imputar el valor mancant.

Respecte el paràmetre m=5, aquest indica que el procediment es realitza 5 vegades per a la imputació de la mateixa variable, i els resultats es guarden. Per defecte però, s'imputarà amb el resultat del primer procediment.

Aquesta tècnica permet substituir els valors mancants per estimacions numèriques basades en els valors de les altres variables presents en el conjunt de dades . La decisió d'optar per la tècnica *MICE* en lloc d'altres s'ha fonamentat en la seguretat i la pràctica d'aquesta tècnica gràcies al coneixement previ.

3.3 Selecció i creació de variables per a la construcció de models predictius

3.3.1 Procés seguit per incorporar noves variables en la base de dades

A continuació, s'explicarà detalladament els passos seguits per obtenir la base de dades necessària a partir del conjunt de dades inicial, amb l'objectiu de realitzar les prediccions finals.

Cal recordar que les variables que es tenen inicialment son les següents:

Data	Últim	Obertura	Màxim	Mínim	Volum	Variació percentual

3.3.1.1 Creació de la variable objectiu

Un pas de rellevant importància en aquest procés és la preparació de la base de dades per a la predicció a tres mesos vista (90 dies). És necessari preparar la base de dades de manera que l'algoritme disposi de la variable objectiu a dia t+90, associada a les característiques corresponents a dia t. Aquest fet es deu a que l'objectiu és, a partir de les variables explicatives a dia t, predir la variable explicada a t+90.

Per tal de fer-ho, es realitzen els següents passos:

1. Per a cada observació, es calcula la data corresponent a 90 dies més enllà de la data original. Per a poder treballar amb aquestes en la fase de modelització, aquesta data es

divideix en tres noves variables: el dia, el mes i l'any, creant així les variables "Dia_t90", "Mes_t90" i "Any_t90".

2. Mitjançant l'ús del llenguatge SQL ("Structured Query Language"), s'ajusta la base de dades per afegir, a cada observació, el valor de la variable "obertura" corresponent a l'observació t+90, creant així una nova variable anomenada "Obertura_90dies", que serà necessària per a crear la variable objectiu, l'explicada.

   Cal destacar que les observacions que cauen en caps de setmana o aquelles que no tenen disponibles el valor d'obertura de t+90 s'eliminen per assegurar la coherència i la integritat del conjunt de dades.

3. Es crea una nova variable que serà la variable objectiu a l'hora d'implementar els models, anomenada "taxa de variació", també coneguda com a taxa de rendibilitat. Aquesta mesura la taxa de variació del preu d'obertura de l'índex a 90 dies i es calcula com la diferència entre l'obertura a 90 dies i l'obertura actual, dividida per l'obertura actual:

   Sent X el preu d'obertura de l'índex :

$$taxa\ de\ variacó = \frac{X(t+90) - X(t)}{X(t)} * 100$$

   Això permet tenir una mesura estandarditzada de rendibilitat que facilita la comparació entre diferents observacions i la construcció dels models predictius.

3.3.1.2 Incorporació del PIB i de la inflació en la base de dades (variables explicatives)

A continuació, s'afegiran dos variables a la base de dades que poden ajudar a determinar el preu dels índexs, per els motius explicats prèviament en el subapartat "Literatura sobre l'anàlisi dels

índexs borsaris i les variables relacionades". Aquestes son el PIB diari de la Eurozona i la inflació diària d'aquesta.

Per a poder incloure el PIB, el que es va realitzar en primer lloc va ser cercar i descarregar les dades necessàries d'una plataforma anomenada "EPDATA" (https://www.epdata.es/variacion-anual-pib-eurozona/533561ad-cf71-4005-a44b-224575b24027) , creada amb l'objectiu de proporcionar informació als periodistes.

Les dades respectives del PIB, es presenten de la següent manera:

Figura 3.9 Mostra de les primeres files de la base de dades corresponent al PIB trimestral

Primeres files de la base de dades del PIB

Año	Periodo	X.
2012	Trimestre 1	-0.5
2012	Trimestre 2	-0.7
2012	Trimestre 3	-1
2012	Trimestre 4	-1
2013	Trimestre 1	-1.1
2013	Trimestre 2	-0.4

Font: Elaboració propia a partir de les dades extretets de EpData

Per tal de poder-les ajuntar amb la base de dades inicial, es crea una nova taula que disposi de dues columnes (mesos de l'any i trimestre associat). Encara falta un últim retoc per tal de poder procedir a ajuntar totes aquestes, que és disposar d'una columna que sigui el mes a data actual (t) i l'any a data actual (t), per així poder associar el PIB diari amb la data corresponent a la base de dades inicial.

Cal fer èmfasis que la base de dades matriu, tal com s'ha explicat, conté les variables "Mes_t90" i "Any_t90", però aquestes no son les dates d'importància, ja que les variables explicatives com el PIB, han de fer referència al moment actual t, no al t+90. Un cop portats a terme els retocs, queden les següents taules:

Figura 3.11 Mostra de les primeres files de la base de dades corresponent al PIB

Primeres files de la base de dades del PIB

Año	Periodo	X.
2012	Trimestre 1	-0.5
2012	Trimestre 2	-0.7
2012	Trimestre 3	-1
2012	Trimestre 4	-1
2013	Trimestre 1	-1.1
2013	Trimestre 2	-0.4

Font: Elaboració propia a partir de les dades extretets de EpData

Figura 3.10 Primeres files de taula relacional entre els mesos i trimestres

Correspondència mes amb trimestre

Periodo	mes
Trimestre 1	1
Trimestre 1	2
Trimestre 1	3
Trimestre 2	4
Trimestre 2	5
Trimestre 2	6

Font: Elaboració propia

Figura 3.12 Primeres files de certes variables de la base de dades matriu

Primeres files de la base de dades

	taxa_rendabilitat	mes_t	any_t
2	8.866708	1	2015
3	11.259349	1	2015
4	11.321449	1	2015
5	6.582433	1	2015
7	6.754662	2	2015
8	6.196734	2	2015

Font: Elaboració pròpia a partir de la base de dades importada de investing.com

Es pot observar que amb les taules presentades ja es pot afegir la variable d'interès, el PIB, a la base de dades principal, i per tant es procedeix a ajuntar la taula 1 amb la dos i la dos amb la tres mitjançant llenguatge SQL.

Un cop finalitzada la inclusió de la variable PIB, es procedeix a afegir la variable inflació. Les dades referents a la inflació s'han obtingut de la pàgina web "EPDATA" , igual que en el cas anterior (https://www.epdata.es/inflacion-zona-euro-ralentiza-sube-16-diciembre-2018/e8ffee44-0684-47e8-b6c2-dde244b6c046) . La variable expressada com a inflació és una mitjana mensual de la inflació de la Eurozona. La diferència amb la variable PIB radica en que la inflació es expressada mensualment, a diferència de PIB que era trimestral. La base de dades descarregada es presenta tal que:

Figura 3.13 Primeres files de la base de dades inflacio_eurozona

Primeres files de la base de dades inflacio_eurozona

Año	Periodo	Inflación.de.la.zona.euro
2005	Enero	1.9
2005	Febrero	2.1
2005	Marzo	2.2
2005	Abril	2.1
2005	Mayo	2
2005	Junio	2

Font: Elaboració propia a partir de les dades extretets de EpData

Per tal d'afegir la variable inflació a la base de dades principal, es crea una funció que s'encarrega de transformar els noms dels mesos a números per així poder-la enllaçar amb les variables "Any_t" i "Mes_t" de la base de dades principal. Un cop realitzades aquestes modificacions, s'afegeix la variable inflació a la base de dades principal, fent que aquesta finalment es presenti de la següent manera:

Figura 3.14 Primeres files de la base de dades resultant

Mostra base de dades amb variables afegides

Data	Ultim	Obertura	Maxim	Minim	Volum	Variacio_percentual	Data_a_90dies	Obertura_90dies	Dia_t90	Mes_t90	Any_t90	taxa_rendabilitat	mes_t	any_t	Inflacio_eurozona	PIB_varperc
2015-01-27	3372.58	3413.33	3413.33	3353.13	66710000	-1.22	2015-04-27	3715.98	27	4	2015	8.866708	1	2015	-0.6	1.7
2015-01-28	3358.96	3384.21	3393.93	3333.36	70370000	-0.40	2015-04-28	3765.25	28	4	2015	11.259349	1	2015	-0.6	1.7
2015-01-29	3371.83	3341.18	3373.18	3326.38	67760000	0.38	2015-04-29	3719.45	29	4	2015	11.321449	1	2015	-0.6	1.7
2015-01-30	3351.44	3386.59	3392.70	3332.39	75030000	-0.60	2015-04-30	3609.51	30	4	2015	6.582433	1	2015	-0.6	1.7
2015-02-03	3414.18	3386.55	3426.67	3386.55	83530000	1.31	2015-05-04	3615.30	4	5	2015	6.754662	2	2015	-0.3	1.7
2015-02-04	3415.53	3416.80	3422.04	3389.20	66530000	0.04	2015-05-05	3628.53	5	5	2015	6.196734	2	2015	-0.3	1.7

Font: Elaboració pròpia a partir de la base de dades importada de investing.com i les variables importades de EpData

3.4 Certs aspectes a tenir en compte previs al modelatge

3.4.1 Identificació de correlacions

Per determinar quines variables són més efectives per construir un model de predicció, cal considerar certs aspectes clau. La inclusió de variables altament correlacionades pot generar un problema de multicol·linealitat, especialment en models lineals. Aquest problema es manifesta quan les variables independents estan altament correlacionades entre elles, cosa que pot afectar la significació individual de cadascuna i fer que els coeficients del model siguin inestables i difícils d'interpretar. Tot i això, és important tenir en compte els següents punts:

En primer lloc, la multicol·linealitat és un problema significatiu quan l'objectiu és contrastar la significació individual de cada variable o assegurar-se del valor exacte de l'efecte d'una variable. Però quan l'objectiu és fer prediccions, com en aquest treball, la multicol·linealitat no és tan preocupant, ja que el focus és en la capacitat predictiva global del model.

En segon lloc, la multicol·linealitat pot ser un problema important en models lineals, però en canvi, en models no lineals com el *KNN*, el random forest i el *GBM*, no és un problema greu. Aquestes metodologies poden gestionar millor les relacions complexes entre variables, incloent-hi aquelles que estan correlacionades.

Tenint en compte això, la millor selecció de variables no ha de dependre de les correlacions entre elles. Per tant, s'optarà per no donar importància a l'existència de correlacions. El procés per seleccionar les millors variables consistirà en provar diferents conjunts de variables per a cada model i avaluar les seves mètriques de rendiment. Aquest procés implica seleccionar un conjunt de variables, construir un model amb aquestes variables, i avaluar el rendiment del model utilitzant quatre mètriques d'error que s'explicaran en el següent subapartat, "3.4.2 Selecció de combinacions de variables en els models".

Segons aquestes mètriques, es decidirà quines variables són més adequades per al model final.

3.4.2 Selecció de combinacions de variables en els models

Tal com s'ha comentat en l'apartat anterior (*"3.4.1 Identificació de correlacions")*, l'elecció de les variables es farà provant diferents conjunts. Concretament, per a obtenir resultats més segurs el que es realitzarà, serà, per a cada model provar totes les combinacions possibles de variables. D'aquesta manera és possible determinar quina combinació funciona millor per a construir el model "guanyador".

El nombre total de combinacions que s'obtindran per a cada algoritme es calcula utilitzant la fórmula de la combinatòria. Per el cas d'aquest treball, amb 10 variables, el nombre total de combinacions serà:

$$\sum_{k=1}^{10} \binom{10}{k} = 2^{10} - 1 = 1023$$

Això significa que tant per al *KNN*, com el random forest com el *GBM*, es provaran 1023 combinacions diferents de variables per trobar aquelles que millor ajustin el model, i per tant, que proporcionin les millors mètriques d'error.

La qüestió és que depenent de la mètrica d'error en la qual es centri, les variables òptimes poden canviar. En el següent subapartat anomenat " *Mètriques d'error per a l'avaluació dels models* ",

s'explicarà detalladament quina mètrica és determina com a l'objectiu de minimització i per quin motiu.

3.4.3 Mètriques d'error per a l'avaluació dels models

Tal com s'ha determinat prèviament , les mètriques d'error que s'avaluaran seran l'error cuadràtic mig, l'arrel de l'error cuadràtic mig, l'error absolut mitjà i l'error percentual absolut mitjà. A continuació, s'explicarà cada una d'aquestes de manera detallada (Contributors to Wikimedia projects, 2024) i es determinarà quina de les quatre es fixarà com a objectiu a minimitzar per tal d'argumentar la selecció del millor model en cada cas.

1. *Mean square error (MSE)* i root *mean square error (RMSE)*

El *Mean Squared Error (MSE)* o error cuadràtic mitg, és una mètrica que tal com diu, mesura l'error mig quadràtic entre els valors predits i els valors reals. Es calcula com la mitjana de les diferències quadrades entre els valors reals i els predits:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Aquest pondera els errors més grans de manera més significativa a causa del quadrat, cosa que pot resultar útil en escenaris on els errors grans s'han de penalitzar més que els errors petits.

El Root *mean square error (RMSE)* o l'arrel de l'error quadràtic mig, és l'arrel quadrada del *MSE*:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

Aquest té la mateixa unitat de mesura que la variable objectiu, cosa que el fa més interpretable. És de rellevant importància, destacar que degut a que el *RMSE* es basa en el *MSE*, hereta la propietat de penalitzar més els errors grans. Això es deu al fet que, en elevar les diferències al quadrat abans de fer la mitjana, els errors grans tenen un impacte gran en el valor final de la mètrica, de la mateixa manera que el *MSE*.

2. *Mean absolute error (MAE)* i *mean absolute percentage error (MAPE)*

El *mean absolute error (MAE)* o error absolut mitjà, calcula la mitjana de les diferències absolutes entre els valors reals i els valors predits:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y_i} |$$

En comparació al *MSE*, el MAE no eleva els errors al quadrat, i això fa que els errors més grans no tinguin tant pes i per tant no penalitza tant els errors de gran mesura tal com ho fa el *MSE i el RMSE*. El MAE, proporciona una mesura directa de l'error mitjà i és més fàcil d'interpretar.

El *mean absolute percentage error (MAPE)* o error percentual absolut mitjà calcula la mitjana de les diferències absolutes percentuals entre els valors reals i els predits:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} | \frac{y_i - \hat{y_i}}{y_i} |$$

Aquest permet avaluar el rendiment del model en termes de la precisió relativa. És important senyalar que el *MAPE* pot tenir problemes amb valors reals propers a zero i pot donar lloc a divisions per zero.

Una vegada definides les diferents mètriques, es determina quina serà la mètrica objecte de minimització per escollir el millor model. En aquest cas, s'ha seleccionat el *RMSE* (Root Mean Squared Error). Tal com s'ha explicat, l'*RMSE* penalitza més fortament els errors grans en comparació amb altres mètriques, cosa que és crucial en l'àmbit de les prediccions financeres, on els errors grans poden tenir conseqüències significatives, com una gran pèrdua monetària. A més, l'*RMSE* és més interpretable que el *MSE*. Donat que l'objectiu d'aquest treball és la predicció de l'índex borsari *Eurostoxx 50*, es considera que la mètrica *RMSE* és la més apropiada.

3.4.4 Validació creuada o cross validation

A continuació s'explicarà el funcionament de la validació creuada, una tècnica de rellevant importància per garantir que les mètriques resultants dels models no siguin fruit de l'aleatorietat de les dades seleccionades. Això assegura el bon funcionament dels models i ajuda a evitar problemes com el sobre ajustament (*overfitting*), el qual s'explicarà amb més detall al següent subapartat, anomenat *"3.4.5 Ntree i overfitting".*

Cal comentar que, per als models construïts amb l'algoritme *KNN*, s'ha decidit implementar la validació creuada de manera automàtica. Això significa que s'ha utilitzat la funció del *KNN* que inclou una opció per realitzar una avaluació creuada automàticament. D'altra banda, per als models construïts amb els algoritmes random forest i GBM, s'ha preferit realitzar la validació creuada de manera manual, per tal de poder entendre millor el funcionament intern d'aquesta I la seva dinàmica.

La validació creuada en el codi programat (implementada manualment) comença amb la divisió del conjunt de dades en cinc subgrups o *folds*, mitjançant la funció *create_folds*, assegurant que cada subgrup contingui una quantitat equitativa d'observacions. Cada *fold* està compost pels índexs de les observacions pertinents a aquest subconjunt, fent referència a la base de dades completa.

Figura 3.15 Llistat de subgrups de la validació creuada.

particions2	list [5]	List of length 5
Fold1	integer [364]	1 8 11 15 23 26 ...
Fold2	integer [363]	3 7 9 12 17 19 ...
Fold3	integer [364]	5 10 13 14 16 20 ...
Fold4	integer [362]	21 25 29 32 44 48 ...
Fold5	integer [363]	2 4 6 18 22 28 ...

Font: Funció de R (aplicada a la base de dades)

Amb els subgrups establerts, es creen vectors per emmagatzemar les mètriques d'avaluació de cada *fold*. Seguidament, s'executa un bucle que itera cinc vegades. En cada iteració, es seleccionen quatre dels cinc grups de dades com a dades d'entrenament i es deixa el grup restant com a dades de prova. Aquest procés es repeteix quatre vegades més, utilitzant cada part com a conjunt de validació una vegada.

En cada iteració, es calculen diversos aspectes. Primerament, es divideixen les dades en dos grups: les dades d'entrenament, que s'utilitzen per ajustar el model, i les dades de prova, que s'usen per validar-lo. A continuació, s'ajusta un model específic per a cada iteració del bucle, i un cop ajustat, el model es fa servir per fer prediccions sobre les dades de prova. Les mètriques d'avaluació del rendiment del model en cada iteració s'emmagatzemen en els vectors corresponents, cosa que permet l'anàlisi del rendiment en cada subgrup de dades. Finalment, es calcula la mitjana de cada mètrica d'avaluació al finalitzar el bucle, proporcionant així una estimació global del rendiment esperat del model en el conjunt complet de dades.

3.4.5 Ntree i overfitting

Un aspecte clau en la construcció de models de random forest i *GBM* és l'elecció òptima del paràmetre *ntree*.

Pel que fa a random forest, segons Breiman (2001): «Use of the Strong Law of Large Numbers shows that they always converge so that *overfitting* is not a problem» (p. 6).

En la tècnica del random forest, cada arbre és un model d'aprenentatge independent que fa prediccions basades en subconjunts diferents de dades i conjunts diferents de característiques. La predicció final de random forest és la mitjana (en regressió) de les prediccions de tots els arbres. Segons *Breiman (2001)*, l'ús de molts arbres en un model de random forest es beneficia de la Llei Forta dels Grans Nombres per dos motius principals. Primerament, la reducció de la variància: a mesura que el nombre d'arbres (ntree) augmenta, la variància de la mitjana de les prediccions dels arbres disminueix i aquesta convergeix al valor esperat de la predicció, fent les prediccions cada vegada més estables i precises. En segon lloc, la resistència al sobreajustament: a diferència d'altres models d'aprenentatge automàtic, com el GBM que poden sobreajustar-se fàcilment a les dades d'entrenament, els random forest són menys propensos a això, ja que les prediccions són el resultat de combinacions d'arbres no correlacionats.

Tot i això, també presenta certes limitacions, com el cost computacional elevat i el fet que, arribats a un cert punt, les millores en precisió no són significatives. Per tant, tal com diu *Breiman (2001)*, augmentar el nombre d'arbres *(ntree)* sempre pot beneficiar el model, però arribarà un punt en què les millores seran insignificants.

Pel que fa a l'algorisme *GBM*, segons *Meir (2022)*: «One of the main concerns with any *machine learning* model is that of *overfitting*. Gradient boosting models are no different, and have a higher probability of *overfitting* as the number of iterations increases».

Tal com exposa *Uri Meir (2022)*, tot i que augmentar el valor de *ntree* pot conduir a millors mètriques d'error, cal tenir en compte que això pot provocar sobreajustament (*overfitting*). El sobreajustament és un concepte crític i delicat en l'entrenament de models, ja que implica que el model s'adapta massa bé a les dades d'entrenament i pot no generalitzar bé a noves dades. Això significa que, tot i obtenir bons resultats en les dades d'entrenament, el model podria no funcionar de manera òptima en dades futures.

En aquest treball, s'ha intentat controlar el risc de sobreajustament mitjançant la validació creuada, ja que, tal com s'ha explicat, aquesta tècnica ajuda a avaluar el rendiment del model en múltiples subconjunts de les dades, proporcionant una estimació més robusta de la seva capacitat de generalització.

Per a la construcció dels models en aquest treball, s'ha decidit provar diferents valors de *ntree* tant per a random forest com per a *GBM*. Els valors de ntree provats per a random forest han estat 100, 200, 500, i 995, mentre que per a *GBM* han estat 100, 500, 995, i 2000. Tot i saber que augmentar el paràmetre ntree generalment redueix la mètrica d'error *RMSE*, per controlar el sobreajustament i fer comparables els dos models finalistes de random forest i GBM, s'ha determinat que ntree=995 és el paràmetre òptim per a tots dos models.

3.5 Modelització de models predictius

3.5.1 Knn

Els models inicials es construiran utilitzant l'algoritme *KNN*. Per seleccionar el model finalista que s'utilitzarà per fer les prediccions amb la metodologia *KNN*, primer s'avaluaran diversos models per identificar i obtenir el millor segons les variables disponibles. A continuació, s'explicarà com s'estructura el codi creat per generar tots els models amb aquesta tècnica i, d'entre ells, seleccionar el més adequat.

En primer lloc, es divideixen les dades en dos conjunts: entrenament i prova. Aquesta divisió es fa de manera aleatòria, destinant el 70% de les dades al conjunt d'entrenament i el 30% restant al conjunt de prova. Això permet entrenar el model amb una part de les dades i avaluar-los amb una altra part diferent.

Seguidament, es defineix un conjunt de variables, que definiran el ventall de variables possibles per a construir els models. El motiu de selecció d'aquestes variables en concret, recau en que son les que es poden considerar més rellevants, deixant de banda les variables "Dia_t90", "Mes_t90" i "Any_t90". El model més extens que s'avaluarà serà el següent:

$$Taxa\ rendibilitat \sim Ultim + Obertura + Maxim + Minim + Volum + Variacio\_percentual + mes\_t + any\_t + Inflacio\_eurozona + PIB\_varperc$$

Per construir i avaluar els diferents models, es desenvolupa una funció anomenada avaluar_model. Aquesta realitza les següents funcions:

1. Selecciona les dades d'entrenament i de prova corresponents a les variables especificades per a cada model (fruit de cada combinació de variables que s'utilitzarà).

2. Entrena un model KNN utilitzant la funció que conté R anomenada *train*, en base al conjunt d'entrenament de dades seleccionat. En aquesta funció, s'han d'especificar diversos paràmetres:

Figura 3.16 Codi que integra la funció "train" a R

```
model <- train(
  x = dades_entrenament_subset,
  y = dades_entrenament$taxa_rendabilitat,
  method = "knn",
  trControl = ctrl,
  preProcess = c("center", "scale"),
  tuneGrid = expand.grid(k = c(1:10))
)
```

Font: Elaboració pròpia

On "x" és el conjunt de dades d'entrenament que correspon a les variables explicatives, mentre que "y" és la variable objectiu, la que es vol predir. L'element "method" especifica quin algoritme de *machine learning* s'està utilitzant per construir el model. "TrControl" fa referència al control per a la validació creuada, ja que:

```
ctrl <- trainControl(method = "cv", number = 5)
```

"PreProcess" implica el preprocessament de dades per centrar-les i escalar-les, necessari en el KNN per a que funcioni de manera òptima. Finalment, "TuneGrid" representa els possibles valors, de 1 a 10, que pot prendre "k" durant l'entrenament del model.

3. Realitza prediccions sobre el conjunt de prova, i les emmagatzema.

4. Calcula el rendiment del model utilitzant diverses mètriques d'error: *MSE (Mean Squared Error)*, *RMSE (Root Mean Squared Error)*, *MAE (Mean Absolute Error) i MAPE (mean absolute percentage error)*.

5. Retorna una llista amb les mètriques d'avaluació i el model entrenat.

Un cop definida la funció *avaluar_model*, es procedeix a aplicar-la a tots els possibles conjunts de variables, creant així un model per a cada combinació. Per tal de realitzar-ho el que es fa s'exposa a continuació.

En primer lloc es crea una llista anomenada resultats que serà la encarregada de guardar per cada model (definit amb diferents combinacions de variables), quines variables s'han utilitzat i quin valor té cada mètrica d'avaluació calculada.

En segon lloc es defineix un bucle en el qual es generen totes les combinacions possibles de les variables utilitzant la funció *combn*. S'ha definit el bucle de tal manera que per cada i (de 1 a 10 variables), es generaran amb la funció *comb* totes aquelles combinacions possibles amb i variables. Com molt bé s'ha explicat anteriorment, es generaran un total de 1023 combinacions:

Figura 3.17 Bucle per a generar totes les combinacions de variables en R

```
for (i in 1:length(variables)) {
  combinacions <- combn(variables, i, simplify = FALSE)
  for (combinacio in combinacions) {
    resultat <- avaluar_model(combinacio)
    resultats[[paste(combinacio, collapse = ", ")]] <- resultat
  }
}
```

Font: Elaboració pròpia

Seguidament, per cada combinació possible de variables generada, s'avalua el model fent ús de la funció *avaluar_model*. Tots els resultats es van guardant en la llista definida prèviament a mesura que el bucle procedeix.

Finalment, es genera un altre bucle per tal de comprovar quina de les combinacions de variables genera un model amb la millor mètrica d'avaluació *RMSE*. Això permet identificar la combinació òptima de variables per a la predicció de la *taxa_rendabilitat,* assegurant que es tria el model amb el millor rendiment.

3.5.2 Boscos aleatoris (random forest)

Per definir el model de *random forest* que s'utilitzarà per fer les prediccions, es seguiran uns passos similars als del model *KNN*. Es valoraran diferents combinacions de variables per identificar i obtenir el millor model basat en el *RMSE*. En aquest cas tal com s'ha comentat, es realitzarà la validació creuada manualment.

En primer lloc, es defineixen els *folds*, que representen el nombre de subgrups en què es dividiran les dades per dur a terme la validació creuada. S'utilitzaran 5 *folds*.

De manera similar a s'ha fet amb el *KNN*, es defineix el conjunt de variables que es podran considerar per a construir els models, sent el el model més extens que s'avaluarà:

$$Taxa\ rendibilitat \sim Ultim + Obertura + Maxim + Minim + Volum + Variacio\_percentual + mes\_t + any\_t + Inflacio\_eurozona + PIB\_varperc$$

Per construir i avaluar els diferents models, es desenvolupa una funció anomenada avaluar_model_rf. Per implementar aquesta funció, cal especificar el paràmetre *ntree* i el vector que conté el conjunt de variables que es vol utilitzar. La funció realitza les següents tasques:

1. Crea la partició dels 5 grups.

2. Defineix vectors per emmagatzemar les mètriques d'error: *MSE, RMSE, MAE* i *MAPE.*

3. Realitza un bucle per avaluar cada fold, on es divideixen les dades en un conjunt d'entrenament i un conjunt de prova.

4. Construeix el model Random Forst, utilitzant les dades d'entrenament del conjunt de variables especificat, amb un paràmetre de *ntree* que s'haurà definit prèviament. S'utilitza la funció randomForest, de la següent manera:

```
model_RF <- randomForest(taxa_rendabilitat ~ ., data = dades_entrenament, ntree = ntree)
```

5. Realitza les prediccions amb el conjunt de prova.

6. Calcula les mètriques d'error, comparant les prediccions amb els valors reals de la variable objectiu taxa_rendabilitat I en retorna la mitjana d'aquestes I el model entrenat.

Un cop definida la funció per avaluar el model, es fa el següent:

En primer lloc es crea una llista anomenada *resultats* per emmagatzemar els resultats de cada combinació de variables.

En segon lloc es genera un bucle que itera sobre totes les combinacions possibles de les variables, repetint el procés que s'ha realitzat per a KNN. Per cada combinació, es crida a la funció avaluar_model_rf i els resultats s'emmagatzemen a la llista resultats.

Finalment genera un altre bucle per tal de comprovar quina de les combinacions de variables genera un model amb les millors mètriques d'avaluació, amb ntree fixat a 100.

Tal com s'ha comentat anteriorment (concretament, en l'apartat *"3.4.5 Ntree i overfitting"*) amb la combinació de variables que defineix el model que ha minimitzat el *RMSE* (amb un *ntree* de valor 100), s'opta per constuir més models amb aquesta combinació de variables, però ara per *ntree=200, ntree=500* i *ntree=995*.

Finalment es seleccionar el millor model, aquell que minimitzi el *RMSE* amb la combinació de variables ja seleccionada.

3.5.3 Gradient boosting machine

Per tal de definir el model finalista que s'utilitzarà per a fer les prediccions (amb la metodologia *GBM*), abans, de la mateixa manera que s'ha fet amb les altres dues metodologies, s'avaluaran diferents models per identificar i obtenir el millor en base a les variables disponibles. Es seguiran uns passos similars als de les dos tècniques anteriors, valorant diferents combinacions de variables per identificar i obtenir el millor model basat en les mètriques d'error definides. En aquest cas, com en el *random forest* es realitzarà la validació creuada manualment.

Concretament, els passos que es seguiran son molt similars als del *random forest*, per tant tot allò explicat de manera més detallada en l'apartat anterior, s'explicarà de manera més general en aquest.

En primer lloc, es defineixen els *subgrups* (5 *folds)*. També, es defineix el conjunt de variables que es consideraran per construir els models, definint aquestes el model més extens que s'avaluarà:

$$Taxa\ rendibilitat \sim Ultim + Obertura + Maxim + Minim + Volum + Variacio\_percentual + mes\_t + any\_t + Inflacio\_eurozona + PIB\_varperc$$

Per construir i avaluar els diferents models, també es desenvolupa una funció anomenada avaluar_model_gbm. Aquesta realitza les següents funcions, sent el punt 4 el diferent respecte *random forest*:

1. Crea la partició dels 5 grups.

2. Defineix vectors per emmagatzemar les mètriques d'error: *MSE, RMSE, MAE i MAPE.*

3. Es realitza un bucle per avaluar cada subgrup, on es divideixen les dades en un conjunt d'entrenament i un conjunt de prova.

4. Es construeix el model *gradient boosting machine (gbm)*, utilitzant les dades d'entrenament i es realitzen les prediccions amb el conjunt de prova. S'utilitza la funció *gbm*, de la següent manera:

Figura 3.18 Codi amb la funció "gbm" a R

```
model_GBM <- gbm(taxa_rendabilitat ~ ., data = dades_entrenament,
                 distribution = "gaussian",
                 n.trees = ntree,
                 interaction.depth = 4)
```

Font: Elaboració pròpia

On, "data" és el conjunt de dades d'entrenament que correspon a les variables explicatives seleccionades, mentre que "distribution" especifica la distribució dels errors en el model, assumint en aquest cas una distribució normal d'aquests. L'element "ntree" fa referència al número d'arbres del model. "Interaction.depth" indica la profunditat de l'arbre, o dit d'altra manera, el número de nivells des de l'arrel fins al node final. Aquesta, s'ha definit com a 4, per tant el número màxim de nivells de decisió que pot tenir l'arbre serà 4.

5. Es calculen les mètriques d'error, comparant les prediccions amb els valors reals de la variable objectiu taxa_rendabilitat i en retorna la mitjana d'aquestes, i el model entrenat.

Un cop definida la funció per avaluar el model, es fa el següent (gairebé igual que en l'algoritme *random forest*).

En primer lloc es crea una llista anomenada *resultats* per emmagatzemar els resultats de cada combinació de variables

Seguidament, es genera un bucle que itera sobre totes les combinacions possibles de les variables, tal com s'ha fet anteriorment. Per cada combinació, es crida a la funció avaluar_model_gbm i els resultats s'emmagatzemen a la llista resultats.

Finalment, es genera un altre bucle per tal de comprovar quina de les combinacions de variables genera un model amb les millors mètriques d'avaluació, amb *ntree* fixat a 100.

Tal com s'ha comentat anteriorment (concretament, en l'apartat *"3.4.5 Ntree i overfitting"*) i amb la combinació de variables que defineix el model que ha minimitzat el *RMSE* , s'opta per constuir més models amb aquesta combinació de variables, però a diferència del random forest, ara per *ntree=500, ntree=995* i *ntree=2000*.

En aquest cas, no es selecciona el millor model segons els criteris establerts, que com ja s'ha expressat es la minimització del *RMSE*, sinó que es selecciona com a finalista aquell per un ntree=995, per tal de que sigui comparable amb el seleccionat com a finalista per al random forest.

4 Resultats

Un cop implementades les funcions i construïts els models, s'obtenen els resultats que s'exposaran a continuació.

4.1 KNN

Respecte *KNN*, per cada mètrica d'avaluació, la millor combinació de variables que s'ha obtingut és la següent:

Taula 4.1 Millors mètriques (MSE, RMSE, MAE, MAPE) per KNN

Mètrica	Millor Combinació	Valor
MSE	Obertura, Maxim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	4.289709
RMSE	Obertura, Maxim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	2.071161
MAE	Obertura, Maxim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	1.508709
MAPE	Obertura, Maxim, Minim, mes_t, any_t	88.777822

Font: Elaboració pròpia a partir de R

A continuació, en base a la taula anterior, es selecciona el model que inclou aquelles variables que minimitzen el *RMSE* . Concretament el RMSE mínim és 2,071161 i el model associat es presenta com:

$$Taxa\ rendibilitat \sim Obertura + Maxim + mes\_t + any\_t + Inflacio\_eurozona + PIB\_varperc$$

El resultat final de totes les mètriques d'avaluació dels error corresponents a aquest model son:

Taula 4.2 Mètriques MSE, RMSE, MAE, MAPE per KNN associades a el mínim RMSE

Mètrica	Valor
MSE	4.289709
RMSE	2.071161
MAE	1.508709
MAPE	89.640636

Font: Elaboració pròpia a partir de R

4.2 Random forest

Respecte la tècnica *random forest*, la millor combinació de variables que s'ha obtingut per cada mètrica d'avaluació és la següent (cal recordar que amb un paràmetre *ntree* fix a 100):

Taula 4.3 Mètriques MSE, RMSE, MAE, MAPE per random forest (ntree=100)

Mètrica	Millor Combinació	Valor
MSE	Obertura, Minim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	4.628379
RMSE	Obertura, Minim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	2.150946
MAE	Obertura, Minim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	1.551576
MAPE	Obertura, Minim, Volum, mes_t, Inflacio_eurozona, PIB_varperc	85.808308

Font: Elaboració pròpia a partir de R

A continuació, en base a la taula anterior, es selecciona el model que inclou aquelles variables que minimitzen el *RMSE* . Concretament el RMSE mínim és 2,150946 i el model associat es presenta com:

$$Taxa\ rendibilitat \sim Obertura + Minim + mes\_t + any\_t + Inflacio\_eurozona + PIB\_varperc$$

Amb el model guanyador definit per les variables presentades, s'observa quin valor tenen les mètriques d'avaluació per els diferents valors de *ntree* prèviament mencionats, que son 200, 500 i 995. Els valors que s'obtenen son els següents:

Taula 4.4 RMSE per a diferents arbres (random forest)

	Nombre d'arbres (ntree)	RMSE
ntree_200	200	2.143757
ntree_500	500	2.134089
ntree_995	995	2.133925

Font: Elaboració pròpia a partir de R

En base a la taula anterior, es selecciona el model amb un paràmetre *ntree* igual a 995, i les mètriques d'aquest es presenten de la següent manera:

Taula 4.5 Mètriques associades a ntree=995 (random forest)

Mètrica	Valor
MSE	4.555690
RMSE	2.133925
MAE	1.541892
MAPE	89.380115

Font: Elaboració pròpia a partir de R

En base a la taula, es pot observar que totes les mètriques associades al model guanyador de random forest, menys la *MAPE*, presenten valors més elevats i per tant pitjors que el les mètriques associades al model seleccionat per *KNN*.

4.3 Gradient boosting machine

Respecte la tècnica *GBM*, la millor combinació de variables que s'ha obtingut per cada mètrica d'avaluació és la següent (cal recordar que amb un paràmetre *ntree* fix a 100):

Taula 4.6 Mètriques MSE, RMSE, MAE, MAPE per GBM (ntree=100)

Mètrica	Millor Combinació	Valor
MSE	Obertura, Minim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	11.100781
RMSE	Obertura, Minim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	3.328412
MAE	Obertura, Minim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	2.475964
MAPE	Obertura, Volum, Variacio_percentual, mes_t, Inflacio_eurozona, PIB_varperc	111.837782

Font: Elaboració pròpia a partir de R

A continuació, en base a la taula anterior, es selecciona el model que inclou aquelles variables que minimitzen el *RMSE* . Concretament el RMSE mínim és 3,328412 i el model associat es presenta com:

$$Taxa\ rendibilitat \sim Obertura + Minim + mes\_t + any\_t + Inflacio\_eurozona + PIB\_varperc$$

Amb el model guanyador definit per les variables presentades, s'observa quin valor tenen les mètriques d'avaluació per els diferents valors de ntree prèviament mencionats, que son 500, 995 i 2000. Els valors que s'obtenen son els següents:

Taula 4.7 RMSE per a diferents arbres (GBM)

	Nombre d'arbres (ntree)	RMSE
ntree_500	500	2.399894
ntree_995	995	2.286862
ntree_2000	2000	2.245745

Font: Elaboració pròpia a partir de R

Es pot observar que amb un paràmetre de ntree definit a 2000 es minimitza el *RMSE*, però tal com s'ha explicat prèviament, s'escollirà el model fixat amb un *ntree* de 995, ja que així és comparable amb el model creat amb l'algoritme random forest. Així doncs, les mètriques del model amb un paràmetre *ntree* de 995 son les següents:

Taula 4.8 Mètriques associades a ntree igual a 995 (GBM)

Mètrica	Valor
MSE	5.244629
RMSE	2.286862
MAE	1.701084
MAPE	103.014286

Font: Elaboració pròpia a partir de R

En base a la taula presentada, s'observa que el *GBM* no només té el *RMSE* més alt, sinó que també presenta totes les altres mètriques més desfavorables en comparació amb els models *KNN* i random forest.

4.4 Comparació de resultats de les tres tècniques usades i selecció del millor model

A continuació, es presenten conjuntament els resultats dels tres models finalistes amb l'objectiu de comparar-los entre ells i posteriorment determinar quin és el més adequat per l'objectiu preestablert. La selecció es basa en la minimització del *RMSE*, tal com s'ha establert anteriorment com a criteri principal.

La següent taula següent mostra les mètriques i les combinacions de variables que formen els models seleccionats per a cada una de les tres tècniques utilitzades:

Taula 4.9 Presentació de les mètriques finals i dels models associats per a les 3 tècniqes de machine learning

Model	Variables Seleccionades	MSE	RMSE	MAE	MAPE
KNN	Obertura, Maxim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	4.289709	2.071161	1.508709	88.77782
Random Forest	Obertura, Minim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	4.628379	2.150946	1.551576	85.80831
GBM	Obertura, Minim, mes_t, any_t, Inflacio_eurozona, PIB_varperc	11.100781	3.328412	2.475964	111.83778

Font: Elaboració pròpia a partir de R

Observant la taula final, es pot observar que el model *KNN* destaca per oferir els millors resultats en termes de minimització de l'error, amb valors de MSE de 4.289709, RMSE de 2.071161 i MAE de 1.508709. El MAPE del KNN, amb un valor de 88.77782, indica una desviació percentual moderada respecte als valors reals.

Un RMSE de 2.071161 indica que les prediccions del model tenen un error mitjà de 2.071161 unitats respecte als valors reals. En termes econòmics, això podria significar que si un inversor es planteja invertir en l'Eurostoxx 50, amb un valor real de 100 unitats, el model KNN podria predir un valor entre 97.928839 i 102.071161, mostrant una precisió acceptable per a decisions d'inversió. Així mateix, un MAE de 1.508709 implica que l'error absolut mitjà és de 1.508709 unitats, suggerint que les prediccions del KNN es mantenen properes als valors reals, fet que augmenta la confiança dels inversors en les seves decisions.

El model random forest presenta un rendiment lleugerament inferior al KNN, amb un MSE de 4.628379, un RMSE de 2.150946 i un MAE de 1.551576. No obstant això, el MAPE per al random forest és de 85.80831, lleugerament millor que el del KNN, la qual cosa indica que les prediccions poden ser més precises percentualment. Aquesta capacitat del random forest per combinar múltiples arbres de decisió ajuda a reduir el risc de sobreajustament i millora la robustesa del model.

El model GBM mostra els resultats menys satisfactoris, amb un MSE de 11.100781, un RMSE de 3.328412 i un MAE de 2.475964. El MAPE de 111.83778 reflecteix una desviació percentual significativa en les prediccions, indicant una baixa precisió.

Un cop analitzats i comparats els resultats obtinguts, es procedeix a seleccionar aquell model que millor compleix l'objectiu de minimització del *RMSE* . Aquest és el pertinent al KNN, amb un RMSE de 2,071, tal com ja s'ha comentat. Finalment, aquest es presenta com:

$$Taxa\ rendibilitat \sim Obertura + Maxim + mes\_t + any\_t + Inflacio\_eurozona + PIB\_varperc$$

5 CONCLUSIONS

Aquest treball demostra que l'ús de tècniques de *machine learning* pot oferir resultats de gran utilitat per als inversors. Entre els models analitzats, el *KNN* destaca com el millor model per a la predicció de la taxa de variació de l'índex Eurostoxx 50 gràcies a la seva capacitat per minimitzar els errors de predicció. Estadísticament, el model KNN demostra una elevada precisió amb una menor desviació respecte als valors reals, fet que el fa ideal per a la predicció d'índexs borsaris, i en aquest cas concretament, per a l'anticipació de la rendibilitat de l'índex Eurostoxx 50.

Des d'una perspectiva econòmica, l'aplicació d'aquest model proporciona informació valuosa per als inversors, permetent una millor gestió del risc i una presa de decisions més informada. Això pot ajudar a complir els objectius determinats, donant veu a aquest tipus d'inversió gràcies a tècniques com les desenvolupades en aquest treball. Aquestes tècniques permeten predir les variacions de l'índex borsari Eurostoxx 50 i així guanyar seguretat en la inversió del fons indexat que reprodueix aquest índex. En termes pràctics, això significa que els inversors poden utilitzar aquest model per avaluar la viabilitat de les seves inversions amb un grau més alt de confiança. Dit d'una altra manera, un inversor propietari d'una cartera d'inversió basada en l'Eurostoxx 50 podria utilitzar el model KNN per obtenir una previsió més precisa de la rendibilitat a tres mesos, minimitzant així el risc associat a les seves decisions d'inversió i decidint d'aquesta manera si li val la pena la seva inversió en base als beneficis mínims que li agradaria obtenir.

No obstant això, és essencial recordar que, malgrat els avenços significatius en el camp del *machine learning* i les prediccions financeres, l'economia i el mercat sempre estan subjectes a la incertesa. Aquesta realitat implica reconèixer les limitacions inherents als models predictius.

Aquest estudi no només subratlla la utilitat de les tècniques de machine learning per a la predicció financera, sinó que també posa de manifest la importància de la selecció adequada de variables i models. La metodologia emprada en aquest treball ofereix un marc sòlid per a futures investigacions i aplicacions pràctiques en l'àmbit de la predicció financera, aportant eines per a una presa de decisions més informada i fonamentada en dades empíriques.

6 DISCUSSIÓ

En aquest apartat es volen destacar alguns punts del treball paral·lels a les conclusions, començant per comentar si els resultats dels models han sigut els que s'esperaven, analitzar les dificultats i les limitacions que s'han trobat al llarg de la seva elaboració, i proposar futures línies d'investigació que podrien millorar els resultats obtinguts i aportar nous coneixements al camp d'estudi.

En primer lloc, comentar que tot i que podria sorprendre que el KNN doni els millors resultats en aquest context, és important destacar que estudis recents han demostrat que, depenent de les dades, models més senzills com el KNN poden ajustar-se millor que models més complexos com el random forest o el GBM. Això es deu a la seva simplicitat i a la menor susceptibilitat a l'*overfitting* en determinades situacions. Per exemple, un estudi publicat a "Sensors" va comparar el rendiment del random forest, KNN i SVM en la classificació de cobertes de sòl i va trobar que el KNN podia competir eficaçment amb els altres models en determinades condicions de dades *(Noi & Kappas, 2017)*. Això és reforçat per altres investigacions que indiquen que el KNN, tot i ser un model més simple, pot ser més estable i oferir millors resultats quan les dades tenen estructures específiques *(Li et al., 2011)*.

Un altre tema de discussió, és la importància de la fiabilitat de les dades. Com molt bé es sap, la qualitat de les dades és fonamental per a l'èxit de qualsevol anàlisi o modelització. En aquest treball, un dels primers reptes va ser assegurar la validesa de les dades utilitzades. Yahoo Finance, una de les fonts més conegudes per a la descàrrega de dades financeres, no permetia la descàrrega de les dades específiques necessàries per a aquest treball. A causa d'aquesta limitació, es van haver d'explorar diferents fonts per obtenir dades fiables, cosa que va retardar l'inici de l'anàlisi. Finalment, es va trobar una altra pàgina anomenada "investing.com", que va permetre la descàrrega de totes les dades necessàries.

Quan es va decidir afegir variables macroeconòmiques com el PIB i la inflació, també van sorgir dificultats, especialment en l'anàlisi temporal d'aquestes. El PIB de la eurozona només es trobava disponible de manera mensual i la inflació de manera trimestral, mentre que les observacions de la base de dades principal eren diàries. Per superar aquest problema i poder utilitzar les dades d'interès, es va decidir replicar les dades mensuals o trimestrals de PIB i inflació per ajustar-les al

format diari de la base de dades matriu. Això va permetre integrar aquestes variables macroeconòmiques en l'anàlisi, tot i les diferències en la freqüència de les dades.

Durant l'elaboració del treball, es va haver de replantejar l'objectiu inicial. En un principi, es pretenia desenvolupar models predictius per a l'S&P 500 i l'*Eurostoxx 50*, amb la finalitat de comparar el rendiment dels dos índexs i determinar quin d'ells era més rendible per invertir a 90 dies vista. Aquest objectiu va haver de ser ajustat a causa de les limitacions imposades per l'extensió habitual dels treballs de final de grau (TFG), ja que incloure el segon índex podia comportar superar la limitació de pàgines establerta. Aquesta situació mostra la importància de ser capaç d'adaptar-se a circumstàncies imprevistes i ajustar els objectius en funció de les condicions. A nivell personal i en relació amb el que s'acaba de comentar, aquest treball ha estat una oportunitat per desenvolupar habilitats d'adaptació a canvis d'organització, un objectiu personal que s'ha estat treballant des de fa temps.

Una altra dificultat significativa trobada durant el treball ha estat la comprensió precisa i detallada del funcionament intern de cada model de *machine learning* utilitzat. Tot i tenir coneixements previs sobre els models KNN, random forest i *gradient boosting machine (GBM)*, es va trobar necessari aprofundir molt més en els detalls tècnics i en els paràmetres de cadascun. En particular, el gradient boosting machine s'ha presentat com el més complex. Aquest model presenta una gran flexibilitat i pot oferir resultats molt bons, però també és susceptible al sobreajustament si els seus paràmetres no es configuren correctament.

Durant el procés de modelització, s'han adquirit molts coneixements sobre les tres metodologies, però encara queda molt per aprendre, especialment pel que fa a l'ús eficient de certs paràmetres. Per exemple, en el cas del GBM, un dels paràmetres clau és el shrinkage, també conegut com a learning rate, que ajuda a controlar el sobreajustament del model. Com a futura línia d'investigació, es proposa incloure aquest paràmetre en els models utilitzats i aprofundir en l'ús d'aquest i altres paràmetres avançats per millorar el rendiment del GBM i garantir la seva capacitat de generalització.

Si es tornés a començar el treball des de zero, una de les millores que es podria implementar seria la cerca de més literatura especialitzada sobre variables determinants dels índexs borsaris que poguessin ajudar a millorar els models predictius. Aquesta proposta de millora es considera una

línia de futura investigació que podria interessar a altres investigadors del camp, ja que aprofundir en la relació entre diferents variables econòmiques i financeres pot proporcionar noves perspectives i eines per a la modelització.

Un altre aspecte que es podria canviar si es reprengués el treball seria l'elecció de les metodologies utilitzades. En lloc de limitar-se a KNN, GBM i random forest, es podria considerar la incorporació de xarxes neuronals, que representen una altra branca important de l'aprenentatge automàtic, concretament l'aprenentatge no supervisat. Les xarxes neuronals tenen un gran potencial per a la modelització predictiva gràcies a la seva capacitat per capturar relacions complexes entre variables. Inicialment, es va plantejar utilitzar xarxes neuronals, però aquesta idea es va descartar a causa de l'extensió del treball que es tenia, havent realitzat els altres models. Per tant, les xarxes neuronals també es proposen com una línia oberta per a futures investigacions, ja que oferirien una visió complementària i podrien millorar encara més els resultats obtinguts.

Una altre possible millora podria estar relacionada en el procés de selecció de variables òptimes per a la construcció dels models. En aquest treball, s'han provat totes les combinacions possibles de variables, emmagatzemant els resultats de les mètriques per identificar la combinació que minimitzava el *RMSE*. Aquest procés, és llarg i molt costós computacionalment. Per tant, es proposa investigar tècniques més eficients i avançades per a la selecció de les millors variables per als models.

7 BIBLIOGRAFIA

Amat Rodrigo, J. (2017, February). *Arboles de decision, Random Forest, Gradient Boosting y C5.0*. https://cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_c50

Arnott, A. (2022, June 16). El verdadero problema de la inflación. *Morningstar, Inc.* https://www.morningstar.es/es/news/224267/el-verdadero-problema-de-la-inflaci%C3%B3n.aspx

Breiman, Leo. (2001a). In *"Random forests", a Machine learning 45.1* (pp. 5–32). https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Breiman, L. (2001b). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, J. (2003, November 23). What are blue chip stocks and are they good investments? *Investopedia*. https://www.investopedia.com/terms/b/bluechipstock.asp

Chen, J. (2006, June 19). Free-Float methodology and how to calculate market capitalization. *Investopedia*. https://www.investopedia.com/terms/f/freefloatmethodology.asp

Chen, T., & Guestrin, C. (2016, August 13). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. http://dx.doi.org/10.1145/2939672.2939785

Contributors to Wikimedia projects. (n.d.). *Capitalització borsària*. Wikipedia. https://ca.wikipedia.org/wiki/Capitalitzaci%C3%B3_bors%C3%A0ria

Contributors to Wikimedia projects. (2022, June 18). *Arbre de decisió*. Wikipedia. https://ca.wikipedia.org/wiki/Arbre_de_decisi%C3%B3

Contributors to Wikimedia projects. (2024a, March 11). *Mean squared error*. Wikipedia. https://en.wikipedia.org/wiki/Mean_squared_error

Contributors to Wikimedia projects. (2024b, April 2). *Mean absolute error*. Wikipedia. https://en.wikipedia.org/wiki/Mean_absolute_error

Contributors to Wikimedia projects. (2024c, May 3). *Root mean square deviation*. Wikipedia. https://en.wikipedia.org/wiki/Root_mean_square_deviation

Contributors to Wikimedia projects. (2024d, May 22). *Mean absolute percentage error*. Wikipedia. https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

Contributors to Wikimedia projects. (2024e, June 10). *Random forest*. Wikipedia. https://ca.wikipedia.org/wiki/Random_forest

Coursera. (2023, January 30). *3 tipos de machine learning que debes conocer*. Coursera. https://www.coursera.org/mx/articles/types-of-machine-learning

Daniel. (2021, December 13). *Machine Learning: Definición, funcionamiento, usos*. Formación En Ciencia de Datos | DataScientest.Com. https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

EpData. (2024, February 31). *Variación anual del PIB de la eurozona*. EpData - La Actualidad Informativa En Datos Estadísticos de Europa Press. https://www.epdata.es/variacion-anual-pib-eurozona/533561ad-cf71-4005-a44b-224575b24027

España, A. (n.d.). *¿Qué es el EuroStoxx 50?* Observatorio Del Inversor. https://www.andbank.es/observatoriodelinversor/que-es-el-eurostoxx-50/

ESPAÑA, B. (n.d.). Activos financieros, qué son y cómo clasificarlos. *BBVA School*. https://www.bbva.es/finanzas-vistazo/ef/fondos-inversion/activos-financieros.html

Eurostoxx 50. (n.d.). DEGIRO. https://www.degiro.es/cotizar/eurostoxx-50

Frost, J. (2018, October 8). *Introduction to bootstrapping in statistics with an example*. Statistics By Jim. https://statisticsbyjim.com/hypothesis-testing/bootstrapping/

GeeksforGeeks. (2017, April 14). K-Nearest neighbor(knn) algorithm. *GeeksforGeeks*. https://www.geeksforgeeks.org/k-nearest-neighbours/

GeeksforGeeks. (2024, February 22). Random forest algorithm in machine learning. *GeeksforGeeks*. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

Giménez Fernández, R., & Cid Zamorano, P. (2014, January 1). *Modelos predictivos de índices bursátiles relevantes para la economía chilena*. Repositorio Académico de La Universidad de Chile . https://repositorio.uchile.cl/handle/2250/117444

Gratton, P. (2003, November 26). What a stock split is and how it works, with an example. *Investopedia*. https://www.investopedia.com/terms/s/stocksplit.asp

Guo, G., Wang, H., Bell, David , & Greer , K. (n.d.). *"KNN Model-Based Approach in Classification", a Lecture Notes in Computer Science, 2888.* https://link-springer-com.sire.ub.edu/chapter/10.1007/978-3-540-39964-3_62

Histórico del Euro Stoxx 50 (STOXX50E) - Investing.com. (n.d.). Investing.Com Español. https://es.investing.com/indices/eu-stoxx50-historical-data

Índice Bursátil ¿Qué es y cuál es su importancia? – Club de Capitales. (2021, April 16). Club de Capitales. https://clubdecapitales.com/educacion/que-es-un-indice-bursatil/

Kennedy, M. (2012, July 28). Differences among price, value, and unweighted indexes. *The Balance*. https://www.thebalancemoney.com/different-types-of-weighted-indexes-1214780

Li, S., Harner, E. J., & Adjeroh, D. A. (2011). Random KNN feature selection - A fast and stable alternative to Random Forests. *BMC Bioinformatics*, *12*(1), 1–11. https://doi.org/10.1186/1471-2105-12-450

Meir, U. (2022, August 28). D.A.R.T — your new weapon against overfitting in boosting models. *Medium*. https://medium.com/@meir412_37692/d-a-r-t-your-new-weapon-against-overfitting-in-boosting-models-9ea4e6aa435b

Noi, P. T., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors*, *18*(1). https://doi.org/10.3390/s18010018

¿Qué es un árbol de decisión? (n.d.). IBM. https://www.ibm.com/es-es/topics/decision-trees

Random forest. (n.d.). Interactive Chaos. https://interactivechaos.com/es/wiki/random-forest

Sagarra García, M. (n.d.). *Apunts "Instruments i mercats financers", Introducció als Mercats Financers*. https://campusvirtual.ub.edu/pluginfile.php/8023644/mod_resource/content/1/01a-Introducci%C3%B3%20als%20Mercats%20Financers.pdf

Staff, C. (2022, August 5). *Machine learning in finance: 10 applications and use cases*. Coursera. https://www.coursera.org/articles/machine-learning-in-finance

STOXX. (n.d.). *EURO STOXX 50®*. STOXX. https://www.stoxx.com/data-index-details?symbol=SX5E

Tuychiev, B. (2023, December 27). A guide to the gradient boosting algorithm. *DataCamp*. https://www.datacamp.com/tutorial/guide-to-the-gradient-boosting-algorithm

Van Buuren, S. (2018). *mice: Multivariate Imputation by Chained Equations*. 68–74.

Zhang, T. (n.d.). *Flow diagram of gradient boosting machine learning method. The ensemble...* ResearchGate. https://www.researchgate.net/figure/Flow-diagram-of-gradient-boosting-machine-learning-method-The-ensemble-classifiers_fig1_351542039

8 CODI

Per a poder accedir a tot el codi creat i a les bases de dades que s'han utilitzat per a l'elaboració d'aques treball, es pot accedir al següent enllaç: https://github.com/claracarner/CODI_TFG.git